

VISIT...

LANZAROTE
Caliente.COM

Online papers on consciousness

Compiled by [David Chalmers](#)

This is a directory of 1123 online papers on consciousness and related topics. Suggestions for addition are welcome. Most papers are by academic philosophers or scientists.

I have now divided this page into three separate pages, as follows:

- Part I: [Philosophy of Consciousness](#) [368 papers]
- Part II: [Other Philosophy of Mind](#) [433 papers]
- Part III: [Science of Consciousness](#) [322 papers]

TABLE OF CONTENTS

I. PHILOSOPHY OF CONSCIOUSNESS

- [The concept of consciousness](#)
- [The explanatory gap](#)
- [Materialism and dualism](#)
- [The knowledge argument](#)
- [Materialism and modality](#)
- [Metaphysics of consciousness](#)
- [Panpsychism](#)
- [Zombies](#)
- [Qualia](#)
- [The contents of consciousness](#)
- [Representationalism](#)
- [Consciousness as higher-order thought](#)
- [Introspection, self-knowledge, and self-consciousness](#)
- [The unity of consciousness](#)
- [The function of consciousness](#)
- [Philosophy of consciousness \(misc.\)](#)

II. OTHER PHILOSOPHY OF MIND

- [Language and thought](#)
- [Propositional attitudes](#)
- [Theories of content](#)

- [Internalism and externalism](#)
- [Concepts](#)
- [Mental content \(misc.\)](#)
- [Folk psychology](#)
- [Metaphysics of mind](#)
- [The self and personal identity](#)
- [Free will](#)
- [Philosophy of mind \(misc.\)](#)
- [The Chinese Room](#)
- [Godel's theorem and AI](#)
- [The Turing test](#)
- [Philosophy of connectionism](#)
- [Philosophy of AI \(misc.\)](#)

III. [SCIENCE OF CONSCIOUSNESS](#)

- [Neuroscience of visual consciousness](#)
- [Consciousness and neuroscience](#)
- [Cognitive models of consciousness](#)
- [Unconscious perception](#)
- [Implicit memory](#)
- [Implicit learning](#)
- [Change blindness and inattention blindness](#)
- [Visual consciousness \(misc.\)](#)
- [Consciousness and psychology \(misc.\)](#)
- [Consciousness in the history of psychology](#)
- [Animal consciousness](#)
- [Consciousness and artificial intelligence](#)
- [Consciousness and physics](#)
- [Phenomenology](#)
- [Miscellaneous](#)

For other sources of online papers, and for bibliographies of around 2000 offline papers on consciousness, see:

- [people with online papers in philosophy](#)
- [web resources related to consciousness, philosophy, and such](#)
- [directory of online philosophy papers](#) (Joe Lau)

Papers on Consciousness (David Chalmers).

This page contains pointers to some papers on consciousness. The papers here should be accessible to people without a background in philosophy. Philosophers might be interested in the more technical papers under [papers on mind, modality, and meaning](#). A separate page has information on my book [The Conscious Mind](#). Most items are in HTML, with a Postscript version also available in some cases. The papers fall into a few distinct categories.

The Problems of Consciousness

[Consciousness and its Place in Nature](#)

This is a new overview paper on the the metaphysics of consciousness. It summarizes arguments against materialism, and uses these to give a detailed taxonomy of reductive and nonreductive views (three each). It covers some of the same ground as the first two papers below (although it's oriented more toward metaphysics than toward science), while also covering some of the more technical material in my book and some new things. It will appear in the forthcoming *Blackwell Guide to Philosophy of Mind*, edited by Stephen Stich and Fritz Warfield.

[Facing Up to the Problem of Consciousness \(PS\)](#)

This paper gives a nontechnical overview of the problems of consciousness and my approach to them. In it I distinguish between the easy problems and the hard problem of consciousness, and argue that the hard problem eludes conventional methods of explanation. I argue that we need a new form of nonreductive explanation, and make some moves toward a detailed nonreductive theory. This paper, based on a talk I gave at the 1994 Tucson conference on consciousness, appeared in a [special issue](#) of the [Journal of Consciousness Studies](#) in 1995, and also in the 1996 collection [Toward a Science of Consciousness](#), edited by Hameroff, Kaszniak, and Scott (MIT Press, 1996).

[Moving Forward on the Problem of Consciousness](#)

After "Facing Up..." was published, about 25 articles commenting on it or on other aspects of the "hard problem" appeared in *JCS* (links to some of these papers are contained in the article). My (lengthy) reply, ["Moving Forward..."](#), appeared in *JCS* vol. 4, pp. 3-46, 1997. All the papers and my reply have been collected in the book, [Explaining Consciousness: The Hard Problem](#) (edited by Jonathan Shear), published by MIT Press in July 1997.

[The Puzzle of Conscious Experience](#)

This paper appeared in *Scientific American* in December 1995. It is essentially an even less technical

version of the first article above, with some pretty pictures. As with most *Scientific American* articles, much of this article was heavily revised by the editors, and there are a few passages that I cringe at. But it's not a bad introduction.

[Consciousness and Cognition \(PS\)](#)

This is an older paper on consciousness, written when I was a graduate student at Indiana. It talks about the odd fact that even if consciousness is not reductively explainable, our *claims* about consciousness should be, and discusses various ways in which this tension might be resolved, eventually proposing a proto-theory of consciousness based on the notions of pattern and information. I no longer agree with everything in this paper, and it gets a bit wild toward the end, but it covers some interesting issues. I've never tried to publish it, but for some reason it is still my favorite among the papers I've written on the subject.

The Science of Consciousness

[On the Search for the Neural Correlate of Consciousness](#)

This is a constructive analysis of the search for the "neural correlate of consciousness" (or the NCC, as it's sometimes called). I argue that because we don't have any way of detecting consciousness directly (i.e., we have no "[consciousness meter](#)"), the search is driven by pre-empirical bridging principles instead. I discuss some of these principles and draw some conclusions about the shape of the search. This paper is largely a transcript of my talk at the 1996 Tucson conference on consciousness, although some fun and games have been omitted (here are some [visuals](#) from the talk). It was published in *Toward a Science of Consciousness II*, edited by Hameroff, Kaszniak, and Scott (MIT Press, 1998).

[What is a Neural Correlate of Consciousness?](#)

This more recent and longer paper deals with some different aspects of the NCC issue, with reference to recent empirical work in the field (e.g. work in visual neuroscience by Logothetis, Milner and Goodale, and others). In particular it addresses what it means to be a neural correlate of consciousness, distinguishes different sorts of NCCs, and discusses the methodology of the search. It raises some questions about the conclusions that can be drawn from lesion studies. This paper was given at the 1998 ASSC conference on Neural Correlates of Consciousness, and appeared in *Neural Correlates of Consciousness: Conceptual and Empirical Questions*, edited by Thomas Metzinger (MIT Press, 2000).

[First-Person Methods in the Science of Consciousness](#)

In this short paper I argue that the task of a science of consciousness is to connect third-person data about brain and behavior to first-person data about conscious experience, and I discuss the difficult question of how we might investigate and represent first-person data. I also discuss some specific issues about

emotion. This paper was written for a Tucson online workshop on emotion and consciousness, and appeared in the Fall 1999 *Consciousness Bulletin* from the Center for Consciousness Studies.

Commentaries and Reviews

[Insentience, Indexicality, and Intensions](#)

This is a commentary on [John Perry's](#) book *[Knowledge, Possibility, and Consciousness](#)* (MIT Press, 2001), which defends a materialist view against a number of arguments (the zombie argument, the knowledge argument, the modal argument), and addresses the discussion in my book. The commentary will appear in a symposium on Perry's book in *Philosophy and Phenomenological Research*

[Availability: The Cognitive Basis of Experience?](#)

Here I argue that the cognitive correlate of conscious experience is *direct availability for global control*, and use this to shed light on a few vexing questions. This was written as a commentary on [Ned Block's](#) paper "[On A Confusion about a Function of Consciousness](#)", and appeared in *[Behavioral and Brain Sciences](#)* and also in the collection *The Nature of Consciousness: Philosophical Debates* from MIT Press, edited by Block, Flanagan, and Guzeldere. This paper overlaps to some extent with "On the search for a neural correlate of consciousness".

[Self-Ascription Without Qualia: A Case-Study](#)

This is a commentary on Alvin Goldman's piece "[The Psychology of Folk Psychology](#)", in *Behavioral and Brain Sciences* (June 1993). The paper contains a zombie thought-experiment or two, for people who like that sort of thing.

[Review of Journal of Consciousness Studies](#)

This is a review of the first issue of the *Journal of Consciousness Studies*. It appeared in the *Times Literary Supplement* in November 1994.

Miscellaneous

[Absent Qualia, Fading Qualia, Dancing Qualia \(PS\)](#)

In this paper I use thought-experiments to argue that functional organization fully determines conscious experience. These thought-experiments involve the gradual replacement of neurons by silicon chips, and similar scenarios. I argue that if "absent qualia" or "inverted qualia" are possible, then phenomena I call "fading qualia" and "dancing qualia" will be possible; but I argue that it is very implausible that fading or

dancing qualia are possible. The resulting position is a sort of "nonreductive functionalism". This paper appears in the collection [Conscious Experience](#) from Ferdinand Schöningh (1995), edited by [Thomas Metzinger](#). (There is also a German version entitled "Fehlende Qualia, schwindende Qualia, tanzende Qualia" (!), but it is not available on the net.)

[What is the Unity of Consciousness?](#)

This paper is co-authored with [Tim Bayne](#). We distinguish a number of different senses in which it might be said that a subject's conscious experiences are unified, and isolate a central notion for which the claim that consciousness is necessarily unified is tenable without being trivial. We then discuss potential counterexamples to this unity thesis, and will consider the implications of the unity thesis for theories of consciousness more generally. This paper is coming out in *The Unity of Consciousness: Binding, Integration, Dissociation* from Oxford University Press, edited by Axel Cleeremans.

Go to:

- [David Chalmers' home page](#)

Papers on Mind, Modality, and Meaning (David Chalmers).

This page contains some papers around the borders of the philosophy of mind and language, metaphysics, and epistemology. These papers (except "[The Extended Mind](#)") mostly presuppose significant philosophical background (non-philosophers might prefer my page of less technical [papers on consciousness](#)). The papers mostly deal with interwoven issues concerning meaning and modality, sometimes against a background of issues in the philosophy of mind. There is some overlap between the papers, as related issues are approached from different angles. Much of this is being cannibalized for a book manuscript currently in progress (working title: *On What Might Be*).

[On Sense and Intension](#)

Over the last few years I have been working on an account of meaning and of possibility on which these notions are closely tied to reason and cognition. The central tool in doing this has been a two-dimensional semantic framework (invoked in many of the papers on this page). This paper serves as an introduction to those ideas, motivated by the defense of a Fregean conception of meaning. In the paper, I articulate some Fregean theses about sense, develop an intensional account of sense on which it is constitutively connected to epistemic possibility, and use this account to deal with various objections to Fregean views. Along the way, the two-dimensional framework as I understand it emerges.

[The Nature of Epistemic Space.](#)

This unpublished paper takes a more foundational approach to some of these issues, grounding some of the key ideas in a notion of epistemic possibility, and an epistemic space made up of maximal epistemic possibilities, or "scenarios". I explore various ways of constructing the space of scenarios -- one tied to centered possible worlds, and one tied directly to epistemic notions. And I outline some applications of the framework from this perspective: e.g. to Fregean sense, narrow content, indicative conditionals, and hyperintensionality.

[The Foundations of Two-Dimensional Semantics](#)

This monster paper (written for the [Barcelona conference](#) on two-dimensionalism) is a sort of "compare-and-contrast" on the various versions of two-dimensional semantics. It starts by motivating this sort of framework, and then discusses in detail the two main sorts of available understandings of the framework: contextual and epistemic understandings. I argue that contextual understandings (e.g. that of Stalnaker) can't do the work that is required, but that an epistemic understanding can. I set out my own understanding in detail, and then locate existing versions of the framework in the conceptual space as set out.

[The Components of Content](#)

This paper tries to do for thought what first paper above does for language: give an account of the contents of thought on which content is closely tied to reason and cognition. I decompose content into epistemic and subjunctive content, both of which are truth-conditional. Epistemic content is generally internal to a cognitive system, and governs rational relations between thoughts, so it can play the role of "narrow" or "cognitive" content. I apply this framework to a number of puzzles (Frege's puzzle, Kripke's puzzle, the problem of the essential indexical, the mode-of-presentation problem, etc.) in the philosophy of mind and the philosophy of language. The unpublished [1995 version](#) of this paper has been fairly widely cited. The new revised version is finally to be published in my anthology [Philosophy of Mind: Classical and Contemporary Readings](#) in 2002.

[Does Conceivability Entail Possibility?](#)

This paper addresses the epistemology of modality, and argues for a sort of modal rationalism (*a priori* access to modality). It distinguishes a number of sorts of conceivability, and with these distinctions in hand argues that certain sorts of conceivability plausibly entail sorts of possibility. The second half of the paper addresses potential gaps between the two, and gives a positive argument for modal rationalism. Lots of interesting issues come up along the way. This paper is forthcoming in *Conceivability and Possibility*, edited by Tamar Gendler and John Hawthorne (Oxford, 2002).

[Conceptual Analysis and Reductive Explanation](#)

This paper, co-authored with Frank Jackson, is a reply to Ned Block and Robert Stalnaker's paper ["Conceptual Analysis, Dualism, and the Explanatory Gap"](#). It doesn't presuppose knowledge of that paper. It defends from first principles the thesis that there is an *a priori* entailment from microphysical and phenomenal truths (plus or minus a bit) to macroscopic truths; it addresses Block and Stalnaker's objections to this thesis; and finally argues that *a priori* entailment is required for reductive explanation and for physicalism. The paper appears in *Philosophical Review* 110:315-61, 2001.

[Materialism and the Metaphysics of Modality](#)

This paper was my response in a symposium on my book *The Conscious Mind* in *Philosophy and Phenomenological Research* in June 1999 (the corresponding [precis](#) of the book is also online). The commentators were Sydney Shoemaker, Brian Loar, Chris Hill & Brian McLaughlin, and Stephen Yablo, all of whom take a "type-B materialist" position on which there is an epistemic gap between physical and phenomenal, but no modal gap. This gets quickly into issues about the 2-D analysis of a posteriori necessity, and whether there are "strong necessities" that escape it. I argue that there are not, and argue for a sort of modal rationalism.

[Mind and Modality](#)

This material (in detailed outline form) corresponds to a series of three lectures I gave at Princeton University in October 1998. The second and third largely contain material since written up in the papers on this page on conceivability and on phenomenal belief. The first lecture contains some material not contained elsewhere.

[The Content and Epistemology of Phenomenal Belief](#)

This long paper has two halves. The first half gives an account of phenomenal concepts and phenomenal beliefs, on which their content is partly constituted by the quality of an experience. The second half applies this account to epistemological issues: e.g. arguing for a limited incorrigibility thesis, defending a sort of foundationalism about phenomenal knowledge, defending the phenomenal realist from certain epistemological problems, and addressing the "Myth of the Given". The paper is forthcoming in *Consciousness: New Philosophical Essays*, edited by Quentin Smith and Alexandr Jokic (Oxford, 2002).

[The Tyranny of the Subjunctive](#)

This is just an extended outline at the moment. I argues for a parallel between indicative and subjunctive conditionals, on the one hand, and the two dimensions of possibility in the 2-D framework. The standard contemporary analysis of possibility and necessity is grounded in subjunctive conditionals. I suggests that this is entirely arbitrary, and has had a distorting effect on many areas of philosophy.

[The Extended Mind](#)

This paper (jointly written with Andy Clark) argues that cognitive states are not necessarily "in the head", for reasons quite independent of the reference-based considerations of Putnam and Burge. Instead, we advocate an "active externalism", focusing on situations in which an organism is coupled with their environment into what is effectively a single cognitive system (as with a person who relies on a notebook as memory, for example). We argue in detail that mental states such as beliefs can be externally constituted in this way. This leads to a reconception of the relation between mind and world. (Reconciling the externalism of this paper with the internalism of "The Components of Content" is left as an exercise for the reader.) This paper was published in *Analysis* 58:10-23, 1998, and was reprinted in *The Philosopher's Annual*, 1998. Here is a [commentary](#).

[Is There Synonymy in Ockham's Mental Language?](#)

This is my sole venture into the history of philosophy so far. It was written when I was a graduate student in Paul Spade's medieval logic class at Indiana. William of Ockham held that we think in a "mental language", not unlike the language of thought that some contemporary philosophers believe in. The question arises whether the mental language can contain synonyms, or whether these are just artifacts of ordinary language. Most people have said no. Here I give some reasons to say yes. This paper is published in *The Cambridge Companion to Ockham*, edited by Spade, published by Cambridge

The Extended Mind

Andy Clark & [David J. Chalmers](#) [*]

Department of Philosophy
Washington University
St. Louis, MO 63130

Department of Philosophy
University of Arizona
Tucson, AZ 85721

andy@twinearth.wustl.edu
chalmers@arizona.edu

*[[Authors are listed in order of degree of belief in the central thesis.]]

[[Published in *Analysis* 58:10-23, 1998. Reprinted in (P. Grim, ed) *The Philosopher's Annual*, vol XXI, 1998.]]

1 Introduction

Where does the mind stop and the rest of the world begin? The question invites two standard replies. Some accept the demarcations of skin and skull, and say that what is outside the body is outside the mind. Others are impressed by arguments suggesting that the meaning of our words "just ain't in the head", and hold that this externalism about meaning carries over into an externalism about mind. We propose to pursue a third position. We advocate a very different sort of externalism: an *active externalism*, based on the active role of the environment in driving cognitive processes.

2 Extended Cognition

Consider three cases of human problem-solving:

(1) A person sits in front of a computer screen which displays images of various two-dimensional geometric shapes and is asked to answer questions concerning the potential fit of such shapes into depicted "sockets". To assess fit, the person must mentally rotate the shapes to align them with the sockets.

(2) A person sits in front of a similar computer screen, but this time can choose either to physically rotate the image on the screen, by pressing a rotate button, or to mentally rotate the image as before. We can also suppose, not unrealistically, that some speed advantage accrues to the physical rotation operation.

(3) Sometime in the cyberpunk future, a person sits in front of a similar computer screen. This agent, however, has the benefit of a neural implant which can perform the rotation operation as fast as the computer in the previous example. The agent must still choose which internal resource to use (the implant or the good old fashioned mental rotation), as each resource makes different demands on attention and other concurrent brain activity.

How much *cognition* is present in these cases? We suggest that all three cases are similar. Case (3) with the neural implant seems clearly to be on a par with case (1). And case (2) with the rotation button displays the same sort of computational structure as case (3), although it is distributed across agent and computer instead of internalized within the agent. If the rotation in case (3) is cognitive, by what right do we count case (2) as fundamentally different? We cannot simply point to the skin/skull boundary as justification, since the legitimacy of that boundary is precisely what is at issue. But nothing else seems different.

The kind of case just described is by no means as exotic as it may at first appear. It is not just the presence of advanced external computing resources which raises the issue, but rather the general tendency of human reasoners to lean heavily on environmental supports. Thus consider the use of pen and paper to perform long multiplication (McClelland et al 1986, Clark 1989), the use of physical rearrangements of letter tiles to prompt word recall in Scrabble (Kirsh 1995), the use of instruments such as the nautical slide rule (Hutchins 1995), and the general paraphernalia of language, books, diagrams, and culture. In all these cases the individual brain performs some operations, while others are delegated to manipulations of external media. Had our brains been different, this distribution of tasks would doubtless have varied.

In fact, even the mental rotation cases described in scenarios (1) and (2) are real. The cases reflect options available to players of the computer game Tetris. In Tetris, falling geometric shapes must be rapidly directed into an appropriate slot in an emerging structure. A rotation button can be used. David Kirsh and Paul Maglio (1994) calculate that the physical rotation of a shape through 90 degrees takes about 100 milliseconds, plus about 200 milliseconds to select the button. To achieve the same result by mental rotation takes about 1000 milliseconds. Kirsh and Maglio go on to present compelling evidence that physical rotation is used not just to position a shape ready to fit a slot, but often to help *determine* whether the shape and the slot are compatible. The latter use constitutes a case of what Kirsh and Maglio call an 'epistemic action'. *Epistemic* actions alter the world so as to aid and augment cognitive processes such as recognition and search. Merely *pragmatic* actions, by contrast, alter the world because some physical change is desirable for its own sake (e.g., putting cement into a hole in a dam).

Epistemic action, we suggest, demands spread of *epistemic credit*. If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process. Cognitive processes ain't (all) in the head!

3 Active Externalism

In these cases, the human organism is linked with an external entity in a two-way interaction, creating a *coupled system* that can be seen as a cognitive system in its own right. All the components in the system play an active causal role, and they jointly govern behavior in the same sort of way that cognition usually does. If we remove the external component the system's behavioral competence will drop, just as it would if we removed part of its brain. Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head.

This externalism differs greatly from standard variety advocated by Putnam (1975) and Burge (1979). When I believe that water is wet and my twin believes that twin water is wet, the external features responsible for the difference in our beliefs are distal and historical, at the other end of a lengthy causal chain. Features of the *present* are not relevant: if I happen to be surrounded by XYZ right now (maybe I have teleported to Twin Earth), my beliefs still concern standard water, because of my history. In these cases, the relevant external features are *passive*. Because of their distal nature, they play no role in driving the cognitive process in the here-and-now. This is reflected by the fact that the actions performed by me and my twin are physically indistinguishable, despite our external differences.

In the cases we describe, by contrast, the relevant external features are *active*, playing a crucial role in the here-and-now. Because they are coupled with the human organism, they have a direct impact on the organism and on its behavior. In these cases, the relevant parts of the world are *in the loop*, not dangling at the other end of a long causal chain. Concentrating on this sort of coupling leads us to an *active externalism*, as opposed to the passive externalism of Putnam and Burge.

Many have complained that even if Putnam and Burge are right about the externality of content, it is not clear that these external aspects play a causal or explanatory role in the generation of action. In counterfactual cases where internal structure is held constant but these external features are changed, behavior looks just the same; so internal structure seems to be doing the crucial work. We will not adjudicate that issue here, but we note that active externalism is not threatened by any such problem. The external features in a coupled system play an ineliminable role - if we retain internal structure but change the external features, behavior may change completely. The external features here are just as causally relevant as typical internal features of the brain.[*]

*[[Much of the appeal of externalism in the philosophy of mind may stem from the intuitive appeal of active externalism. Externalists often make analogies involving external features in coupled systems, and appeal to the arbitrariness of boundaries between brain and environment. But these intuitions sit uneasily with the letter of standard externalism. In most of the Putnam/Burge cases, the immediate environment is irrelevant; only the historical environment counts. Debate has focused on the question of whether mind must be in the head, but a more relevant question in assessing these examples might be: is mind in the present?]]

By embracing an active externalism, we allow a more natural explanation of all sorts of actions. Once can explain my choice of words in Scrabble, for example, as the outcome of an extended cognitive process involving the rearrangement of tiles on my tray. Of course, one could always try to explain my

action in terms of internal processes and a long series of "inputs" and "actions", but this explanation would be needless complex. If an isomorphic process were going on in the head, we would feel no urge to characterize it in this cumbersome way.[*] In a very real sense, the re-arrangement of tiles on the tray is not part of action; it is part of *thought*.

*[[Herbert Simon (1981) once suggested that we view internal memory as, in effect, an external resource upon which "real" inner processes operate. "Search in memory," he comments, "is not very different from search of the external environment." Simon's view at least has the virtue of treating internal and external processing with the parity they deserve, but we suspect that on his view the mind will shrink too small for most people's tastes.]]

The view we advocate here is reflected by a growing body of research in cognitive science. In areas as diverse as the theory of situated cognition (Suchman 1987), studies of real-world-robotics (Beer 1989), dynamical approaches to child development (Thelen and Smith 1994), and research on the cognitive properties of collectives of agents (Hutchins 1995), cognition is often taken to be continuous with processes in the environment.[*] Thus, in seeing cognition as extended one is not merely making a terminological decision; it makes a significant difference to the methodology of scientific investigation. In effect, explanatory methods that might once have been thought appropriate only for the analysis of "inner" processes are now being adapted for the study of the outer, and there is promise that our understanding of cognition will become richer for it.

*[[Philosophical views of a similar spirit can be found in Haugeland 1995, McClamrock 1985, Varela et al 1991, and Wilson 1994..]]

Some find this sort of externalism unpalatable. One reason may be that many identify the cognitive with the conscious, and it seems far from plausible that consciousness extends outside the head in these cases. But not every cognitive process, at least on standard usage, is a conscious process. It is widely accepted that all sorts of processes beyond the borders of consciousness play a crucial role in cognitive processing: in the retrieval of memories, linguistic processes, and skill acquisition, for example. So the mere fact that external processes are external where consciousness is internal is no reason to deny that those processes are cognitive.

More interestingly, one might argue that what keeps real cognition processes in the head is the requirement that cognitive processes be *portable*. Here, we are moved by a vision of what might be called the Naked Mind: a package of resources and operations we can always bring to bear on a cognitive task, regardless of the local environment. On this view, the trouble with coupled systems is that they are too easily *decoupled*. The true cognitive processes are those that lie at the constant core of the system; anything else is an add-on extra.

There is something to this objection. The brain (or brain and body) comprises a package of basic, portable, cognitive resources that is of interest in its own right. These resources may incorporate bodily actions into cognitive processes, as when we use our fingers as working memory in a tricky calculation, but they will not encompass the more contingent aspects of our external environment, such as a pocket calculator. Still, mere contingency of coupling does not rule out cognitive status. In the distant future we

may be able to plug various modules into our brain to help us out: a module for extra short-term memory when we need it, for example. When a module is plugged in, the processes involving it are just as cognitive as if they had been there all along.[*]

*[[Or consider the following passage from a recent science fiction novel (McHugh 1992, p. 213): "I am taken to the system's department where I am attuned to the system. All I do is jack in and then a technician instructs the system to attune and it does. I jack out and query the time. 10:52. The information pops up. Always before I could only access information when I was jacked in, it gave me a sense that I knew what I thought and what the system told me, but now, how do I know what is system and what is Zhang?"]]

Even if one were to make the portability criterion pivotal, active externalism would not be undermined. Counting on our fingers has already been let in the door, for example, and it is easy to push things further. Think of the old image of the engineer with a slide rule hanging from his belt wherever he goes. What if people always carried a pocket calculator, or had them implanted? The real moral of the portability intuition is that for coupled systems to be relevant to the core of cognition, *reliable* coupling is required. It happens that most reliable coupling takes place within the brain, but there can easily be reliable coupling with the environment as well. If the resources of my calculator or my Filofax are always there when I need them, then they are coupled with me as reliably as we need. In effect, they are part of the basic package of cognitive resources that I bring to bear on the everyday world. These systems cannot be impugned simply on the basis of the danger of discrete damage, loss, or malfunction, or because of any occasional decoupling: the biological brain is in similar danger, and occasionally loses capacities temporarily in episodes of sleep, intoxication, and emotion. If the relevant capacities are generally there when they are required, this is coupling enough.

Moreover, it may be that the biological brain has in fact evolved and matured in ways which factor in the reliable presence of a manipulable external environment. It certainly seems that evolution has favored on-board capacities which are especially geared to parasitizing the local environment so as to reduce memory load, and even to transform the nature of the computational problems themselves. Our visual systems have evolved to rely on their environment in various ways: they exploit contingent facts about the structure of natural scenes (e.g. Ullman and Richards 1984), for example, and they take advantage of the computational shortcuts afforded by bodily motion and locomotion (e.g. Blake and Yuille, 1992). Perhaps there are other cases where evolution has found it advantageous to exploit the possibility of the environment being in the cognitive loop. If so, then external coupling is part of the truly basic package of cognitive resources that we bring to bear on the world.

Language may be an example. Language appears to be a central means by which cognitive processes are extended into the world. Think of a group of people brainstorming around a table, or a philosopher who thinks best by writing, developing her ideas as she goes. It may be that language evolved, in part, to enable such extensions of our cognitive resources within actively coupled systems.

Within the lifetime of an organism, too, individual learning may have molded the brain in ways that rely on cognitive extensions that surrounded us as we learned. Language is again a central example here, as are the various physical and computational artifacts that are routinely used as cognitive extensions by

children in schools and by trainees in numerous professions. In such cases the brain develops in a way that complements the external structures, and learns to play its role within a unified, densely coupled system. Once we recognize that the crucial role of the environment in constraining the evolution and development of cognition, we see that extended cognition is a core cognitive process, not an add-on extra.

An analogy may be helpful. The extraordinary efficiency of the fish as a swimming device is partly due, it now seems, to an evolved capacity to couple its swimming behaviors to the pools of external kinetic energy found as swirls, eddies and vortices in its watery environment (see Triantafyllou and G. Triantafyllou 1995). These vortices include both naturally occurring ones (e.g., where water hits a rock) and self-induced ones (created by well-timed tail flaps). The fish swims by building these externally occurring processes into the very heart of its locomotion routines. The fish and surrounding vortices together constitute a reliable feature of the human environment, such as the sea of words. This linguistic surround envelopes us from birth. Under such conditions, the plastic human brain will surely come to treat such structures as a reliable resource to be factored into the shaping of on-board cognitive routines. Where the fish flaps its tail to set up the eddies and vortices it subsequently exploits, we intervene in multiple linguistic media, creating local structures and disturbances whose reliable presence drives our ongoing internal processes. Words and external symbols are thus paramount among the cognitive vortices which help constitute human thought.

4 From Cognition to Mind

So far we have spoken largely about "cognitive processing", and argued for its extension into the environment. Some might think that the conclusion has been bought too cheaply. Perhaps some *processing* takes place in the environment, but what of *mind*? Everything we have said so far is compatible with the view that truly mental states - experiences, beliefs, desires, emotions, and so on - are all determined by states of the brain. Perhaps what is truly mental is internal, after all?

We propose to take things a step further. While some mental states, such as experiences, may be determined internally, there are other cases in which external factors make a significant contribution. In particular, we will argue that *beliefs* can be constituted partly by features of the environment, when those features play the right sort of role in driving cognitive processes. If so, the mind extends into the world.

First, consider a normal case of belief embedded in memory. Inga hears from a friend that there is an exhibition at the Museum of Modern Art, and decides to go see it. She thinks for a moment and recalls that the museum is on 53rd Street, so she walks to 53rd Street and goes into the museum. It seems clear that Inga believes that the museum is on 53rd Street, and that she believed this even before she consulted her memory. It was not previously an *occurrent* belief, but then neither are most of our beliefs. The belief was sitting somewhere in memory, waiting to be accessed.

Now consider Otto. Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him

everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. Today, Otto hears about the exhibition at the Museum of Modern Art, and decides to go see it. He consults the notebook, which says that the museum is on 53rd Street, so he walks to 53rd Street and goes into the museum.

Clearly, Otto walked to 53rd Street because he wanted to go to the museum and he believed the museum was on 53rd Street. And just as Inga had her belief even before she consulted her memory, it seems reasonable to say that Otto believed the museum was on 53rd Street even before consulting his notebook. For in relevant respects the cases are entirely analogous: the notebook plays for Otto the same role that memory plays for Inga. The information in the notebook functions just like the information constituting an ordinary non-occurrent belief; it just happens that this information lies beyond the skin.

The alternative is to say that Otto has no belief about the matter until he consults his notebook; at best, he believes that the museum is located at the address in the notebook. But if we follow Otto around for a while, we will see how unnatural this way of speaking is. Otto is constantly using his notebook as a matter of course. It is central to his actions in all sorts of contexts, in the way that an ordinary memory is central in an ordinary life. The same information might come up again and again, perhaps being slightly modified on occasion, before retreating into the recesses of his artificial memory. To say that the beliefs disappear when the notebook is filed away seems to miss the big picture in just the same way as saying that Inga's beliefs disappear as soon as she is longer conscious of them. In both cases the information is reliably there when needed, available to consciousness and available to guide action, in just the way that we expect a belief to be.

Certainly, insofar as beliefs and desires are characterized by their explanatory roles, Otto's and Inga's cases seem to be on a par: the essential causal dynamics of the two cases mirror each other precisely. We are happy to explain Inga's action in terms of her occurrent desire to go to the museum and her standing belief that the museum is on 53rd street, and we should be happy to explain Otto's action in the same way. The alternative is to explain Otto's action in terms of his occurrent desire to go to the museum, his standing belief that the Museum is on the location written in the notebook, and the accessible fact that the notebook says the Museum is on 53rd Street; but this complicates the explanation unnecessarily. If we must resort to explaining Otto's action this way, then we must also do so for the countless other actions in which his notebook is involved; in each of the explanations, there will be an extra term involving the notebook. We submit that to explain things this way is to take *one step too many*. It is pointlessly complex, in the same way that it would be pointlessly complex to explain Inga's actions in terms of beliefs about her memory. The notebook is a constant for Otto, in the same way that memory is a constant for Inga; to point to it in every belief/desire explanation would be redundant. In an explanation, simplicity is power.

If this is right, we can even construct the case of Twin Otto, who is just like Otto except that a while ago he mistakenly wrote in his notebook that the Museum of Modern Art was on 51st Street. Today, Twin Otto is a physical duplicate of Otto from the skin in, but his notebook differs. Consequently, Twin Otto is best characterized as believing that the museum is on 51st Street, where Otto believes it is on 53rd. In

these cases, a belief is simply not in the head.

This mirrors the conclusion of Putnam and Burge, but again there are important differences. In the Putnam/Burge cases, the external features constituting differences in belief are distal and historical, so that twins in these cases produce physically indistinguishable behavior. In the cases we are describing, the relevant external features play an active role in the here-and-now, and have a direct impact on behavior. Where Otto walks to 53rd Street, Twin Otto walks to 51st. There is no question of explanatory irrelevance for this sort of external belief content; it is introduced precisely because of the central explanatory role that it plays. Like the Putnam and Burge cases, these cases involve differences in reference and truth-conditions, but they also involve differences in the dynamics of *cognition*.[*]

*[[In the terminology of Chalmers' "The Components of Content" (forthcoming): the twins in the Putnam and Burge cases differ only in their *relational* content, but Otto and his twin can be seen to differ in their *notional* content, which is the sort of content that governs cognition. Notional content is generally internal to a cognitive system, but in this case the cognitive system is itself effectively extended to include the notebook.]]

The moral is that when it comes to belief, there is nothing sacred about skull and skin. What makes some information count as a belief is the role it plays, and there is no reason why the relevant role can be played only from inside the body.

Some will resist this conclusion. An opponent might put her foot down and insist that as she uses the term "belief", or perhaps even according to standard usage, Otto simply does not qualify as believing that the museum is on 53rd Street. We do not intend to debate what is standard usage; our broader point is that the notion of belief *ought* to be used so that Otto qualifies as having the belief in question. In all *important* respects, Otto's case is similar to a standard case of (non-occurrent) belief. The differences between Otto's case and Inga's are striking, but they are superficial. By using the "belief" notion in a wider way, it picks out something more akin to a natural kind. The notion becomes deeper and more unified, and is more useful in explanation.

To provide substantial resistance, an opponent has to show that Otto's and Inga's cases differ in some important and relevant respect. But in what deep respect are the cases different? To make the case *solely* on the grounds that information is in the head in one case but not in the other would be to beg the question. If this difference is relevant to a difference in belief, it is surely not *primitively* relevant. To justify the different treatment, we must find some more basic underlying difference between the two.

It might be suggested that the cases are relevantly different in that Inga has more *reliable* access to the information. After all, someone might take away Otto's notebook at any time, but Inga's memory is safer. It is not implausible that constancy is relevant: indeed, the fact that Otto always uses his notebook played some role in our justifying its cognitive status. If Otto were consulting a guidebook as a one-off, we would be much less likely to ascribe him a standing belief. But in the original case, Otto's access to the notebook is very reliable - not perfectly reliable, to be sure, but then neither is Inga's access to her memory. A surgeon might tamper with her brain, or more mundanely, she might have too much to drink. The mere possibility of such tampering is not enough to deny her the belief.

Perhaps the intuition that Otto's is not a true belief comes from a residual feeling that the only true beliefs are occurrent beliefs. If we take this feeling seriously, Inga's belief will be ruled out too, as will many beliefs that we attribute in everyday life. This would be an extreme view, but it may be the most consistent way to deny Otto's belief. Upon even a slightly less extreme view - the view that a belief must be *available* for consciousness, for example - Otto's notebook entry seems to qualify just as well as Inga's memory. Once dispositional beliefs are let in the door, it is difficult to resist the conclusion that Otto's notebook has all the relevant dispositions.

5 Beyond the Outer Limits

If the thesis is accepted, how far should we go? All sorts of puzzle cases spring to mind. What of the amnesic villagers in *100 Years of Solitude*, who forget the names for everything and so hang labels everywhere? Does the information in my Filofax count as part of my memory? If Otto's notebook has been tampered with, does he believe the newly-installed information? Do I believe the contents of the page in front of me before I read it? Is my cognitive state somehow spread across the Internet?

We do not think that there are categorical answers to all of these questions, and we will not give them. But to help understand what is involved in ascriptions of extended belief, we can at least examine the features of our central case that make the notion so clearly applicable there. First, the notebook is a constant in Otto's life - in cases where the information in the notebook would be relevant, he will rarely take action without consulting it. Second, the information in the notebook is directly available without difficulty. Third, upon retrieving information from the notebook he automatically endorses it. Fourth, the information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequence of this endorsement.[*] The status of the fourth feature as a criterion for belief is arguable (perhaps one can acquire beliefs through subliminal perception, or through memory tampering?), but the first three features certainly play a crucial role.

*[[The constancy and past-endorsement criteria may suggest that history is partly constitutive of belief. One might react to this by removing any historical component (giving a purely dispositional reading of the constancy criterion and eliminating the past-endorsement criterion, for example), or one might allow such a component as long as the main burden is carried by features of the present.]]

Insofar as increasingly exotic puzzle cases lack these features, the applicability of the notion of "belief" gradually falls of. If I rarely take relevant action without consulting my Filofax, for example, its status within my cognitive system will resemble that of the notebook in Otto's. But if I often act without consultation - for example, if I sometimes answer relevant questions with "I don't know" - then information in it counts less clearly as part of my belief system. The Internet is likely to fail on multiple counts, unless I am unusually computer-reliant, facile with the technology, and trusting, but information in certain files on my computer may qualify. In intermediate cases, the question of whether a belief is present may be indeterminate, or the answer may depend on the varying standards that are at play in various contexts in which the question might be asked. But any indeterminacy here does not mean that in the central cases, the answer is not clear.

One might worry that Otto's access to his notebook *in fact* comes and goes. He showers without the notebook, for example, and he cannot read it when it is dark. Surely his belief cannot come and go so easily? We could get around this problem by redescribing the situation, but in any case an occasional temporary disconnection does not threaten our claim. After all, when Inga is asleep, or when she is intoxicated, we do not say that her belief disappears. What really counts is that the information is easily available when the subject needs it, and this constraint is satisfied equally in the two cases. If Otto's notebook were often unavailable to him at times when the information in it would be useful, there might be a problem, as the information would not be able to play the action-guiding role that is central to belief; but if it is easily available in most relevant situations, the belief is not endangered.

Perhaps a difference is that Inga has *better* access to the information than Otto does? Inga's "central" processes and her memory probably have a relatively high-bandwidth link between them, compared to the low-grade connection between Otto and his notebook. But this alone does not make a difference between believing and not believing. Consider Inga's museum-going friend Lucy, whose biological memory has only a low-grade link to her central systems, due to nonstandard biology or past misadventures. Processing in Lucy's case might be less efficient, but as long as the relevant information is accessible, Lucy clearly believes that the museum is on 53rd Street. If the connection was too indirect - if Lucy had to struggle hard to retrieve the information with mixed results, or a psychotherapist's aid were needed - we might become more reluctant to ascribe the belief, but such cases are well beyond Otto's situation, in which the information is easily accessible.

Another suggestion could be that Otto has access to the relevant information only by *perception*, whereas Inga has more direct access -- by introspection, perhaps. In some ways, however, to put things this way is to beg the question. After all, we are in effect advocating a point of view on which Otto's internal processes and his notebook constitute a single cognitive system. From the standpoint of this system, the flow of information between notebook and brain is not perceptual at all; it does not involve the impact of something outside the system. It is more akin to information flow within the brain. The only deep way in which the access is perceptual is that in Otto's case, there is a distinctly perceptual phenomenology associated with the retrieval of the information, whereas in Inga's case there is not. But why should the nature of an associated phenomenology make a difference to the status of a belief? Inga's memory may have some associated phenomenology, but it is still a belief. The phenomenology is not visual, to be sure. But for visual phenomenology consider the Terminator, from the Arnold Schwarzenegger movie of the same name. When he recalls some information from memory, it is "displayed" before him in his visual field (presumably he is conscious of it, as there are frequent shots depicting his point of view). The fact that standing memories are recalled in this unusual way surely makes little difference to their status as standing beliefs.

These various small differences between Otto's and Inga's cases are all *shallow* differences. To focus on them would be to miss the way in which for Otto, notebook entries play just the sort of role that beliefs play in guiding most people's lives.

What about socially extended cognition? Could my mental states be partly constituted by the states of other thinkers? We see no reason why not, in principle. In an unusually interdependent couple, it is entirely possible that one partner's beliefs will play the same sort of role for the other as the notebook plays for Otto.[*] What is central is a high degree of trust, reliance, and accessibility. In other social relationships these criteria may not be so clearly fulfilled, but they might nevertheless be fulfilled in specific domains. For example, the waiter at my favorite restaurant might act as a repository of my beliefs about my favorite meals (this might even be construed as a case of extended desire). In other cases, one's beliefs might be embodied in one's secretary, one's accountant, or one's collaborator.[*]

*[[From the *New York Times*, March 30, 1995, p.B7, in an article on former UCLA basketball coach John Wooden: "Wooden and his wife attended 36 straight Final Fours, and she invariably served as his memory bank. Nell Wooden rarely forgot a name - her husband rarely remembered one - and in the standing-room-only Final Four lobbies, she would recognize people for him."]]

*[[Might this sort of reasoning also allow something like Burge's extended "arthritis" beliefs? After all, I might always defer to my doctor in taking relevant actions concerning my disease. Perhaps so, but there are some clear differences. For example, any extended beliefs would be grounded in an existing active relationship with the doctor, rather than in a historical relationship to a language community. And on the current analysis, my deference to the doctor would tend to yield something like a true belief that I have some other disease in my thigh, rather than the false belief that I have arthritis there. On the other hand, if I used medical experts solely as terminological consultants, the results of Burge's analysis might be mirrored.]]

In each of these cases, the major burden of the coupling between agents is carried by language. Without language, we might be much more akin to discrete Cartesian "inner" minds, in which high-level cognition relies largely on internal resources. But the advent of language has allowed us to spread this burden into the world. Language, thus construed, is not a mirror of our inner states but a complement to them. It serves as a tool whose role is to extend cognition in ways that on-board devices cannot. Indeed, it may be that the intellectual explosion in recent evolutionary time is due as much to this linguistically-enabled extension of cognition as to any independent development in our inner cognitive resources.

What, finally, of the self? Does the extended mind imply an extended self? It seems so. Most of us already accept that the self outstrips the boundaries of consciousness; my dispositional beliefs, for example, constitute in some deep sense part of who I am. If so, then these boundaries may also fall beyond the skin. The information in Otto's notebook, for example, is a central part of his identity as a cognitive agent. What this comes to is that Otto *himself* is best regarded as an extended system, a coupling of biological organism and external resources. To consistently resist this conclusion, we would have to shrink the self into a mere bundle of occurrent states, severely threatening its deep psychological continuity. Far better to take the broader view, and see agents themselves as spread into the world.

As with any reconception of ourselves, this view will have significant consequences. There are obvious consequences for philosophical views of the mind and for the methodology of research in cognitive science, but there will also be effects in the moral and social domains. It may be, for example, that in some cases interfering with someone's environment will have the same moral significance as interfering

with their person. And if the view is taken seriously, certain forms of social activity might be reconceived as less akin to communication and action, and as more akin to thought. In any case, once the hegemony of skin and skull is usurped, we may be able to see ourselves more truly as creatures of the world.

REFERENCES

Beer, R. 1989. *Intelligence as Adaptive Behavior*. New York: Academic Press.

Blake, A. & Yuille, A. (eds) 1992. *Active Vision*. Cambridge, MA: MIT Press.

Burge, T. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4:73-122.

Clark, A. 1989. *Microcognition*. MIT Press.

Haugeland, J. 1995. Mind embodied and embedded. In (Y. Houng and J. Ho, eds.), *Mind and Cognition*. Taipei: Academia Sinica.

Hutchins, E. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press,

Kirsh, D. 1995. The intelligent use of space. *Artificial Intelligence* 73:31-68.

Kirsh, D. & Maglio, P. 1994. On distinguishing epistemic from pragmatic action. *Cognitive Science* 18:513-49.

McClamrock, R. 1995. *Existential Cognition*. Chicago: University of Chicago Press.

McClelland, J.L, D.E. Rumelhart, & G.E. Hinton 1986. The appeal of parallel distributed processing". In (McClelland & Rumelhart, eds) *Parallel Distributed Processing, Volume 2*. Cambridge, MA: MIT Press.

McHugh, M. 1992. *China Mountain Zhang*. New York: Tom Doherty Associates.

Putnam, H. 1975. The meaning of `meaning'. In (K. Gunderson, ed) *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press.

Simon, H. 1981. *The Sciences of the Artificial*. MIT Press.

Suchman, L. 1987. *Plans and Situated Actions*. Cambridge, UK: Cambridge University Press.

Thelen, E. & Smith, L. 1994. *A Dynamic Systems Approach to the Development of Cognition and Action*.

On Sense and Intension

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

What is involved in the meaning of our expressions? Frege suggested that there is an aspect of an expression's meaning - the expression's *sense* - that is constitutively tied to the expression's role in reason and cognition. Many contemporary philosophers have argued that there is no such aspect of meaning. I think that Frege was closer to the truth: one can make good sense of an aspect of meaning with many, although not all, of the properties that Frege attributed to sense. One can naturally associate an expression with an *epistemic intension* - a function from epistemic possibilities to extensions - such that an expression's epistemic intension is deeply tied to its role in reason and cognition, and such that it has many of the properties attributed to Fregean sense. I will argue that the claim that expressions have associated intensions of this sort is undefeated by the major contemporary arguments against Fregean sense.

1 Introduction

The simplest aspect of an expression's meaning is its *extension*. We can stipulate that the extension of a sentence is its truth-value, and that the extension of a singular term is its referent. The extension of other expressions can be seen as associated entities that contribute to the truth-value of a sentence in a manner broadly analogous to the way in which the referent of a singular term contributes to the truth-value of a sentence. In many cases, the extension of an expression will be what we intuitively think of as its referent, although this need not hold in all cases, as the case of sentences illustrates. While Frege himself is often interpreted as holding that a sentence's truth-value is its referent, this claim is counterintuitive and widely disputed. We can avoid that issue in the present framework by using the technical term 'extension'; in this context, the claim that the extension of a sentence is its truth-value is a stipulation.

Different sorts of expressions have different sorts of extensions. By the stipulation above, the extension of a singular term is an individual: the extension of 'France' is a particular country (France), and the extension of 'Bill Clinton' is a particular person (Clinton). Analogously, the extension of a general term is plausibly seen as a class: the extension of 'cat' is a particular class of animals (the class of cats). The extension of a kind term can be seen as a kind: the extension of 'water' is a particular substance (water). The extension of a predicate can be seen as a property or perhaps as a class: the extension of 'hot' is a particular property (hotness) or a particular class (the class of hot things).

The extensions of expressions can at least often be used to determine the extensions of complex expressions of which they are parts. This applies most obviously to the truth-values of sentences. For example, 'Sydney is in Australia' is true, and it is true because the extension of 'Sydney' (a particular city) is located in the extension of 'Australia' (a particular country). 'Michael Jordan is short' is false, and it is false because the individual who is the extension of 'Michael Jordan' does not have the property that is the extension of 'short'. This also applies to complex expressions shorter than sentences: for example, it is not implausible that the complex singular term 'the greatest cricket player' has an extension (Don Bradman), and that this extension is determined by the extensions of its parts.

There are various complexities here, and there are corresponding choices to be made. For example, some terms (e.g. 'Santa Claus') appear to have no referent: in such a case, one might say that they lack extension, or one might say that they have a null extension. In some cases (e.g. 'greatest' above), it appears that the extension of an expression can depend on context: for this reason, we may wish to assign extensions to expression tokens, or to expression types in contexts, rather than to expression types alone. In general, the truth-value of a sentence will be determined by the extensions of its parts within a regimented *logical form*, along with corresponding principles for determining truth-value of a regimented sentence from its logical form and the extensions of its parts; and the regimented sentence may look quite different from the original sentence, with different basic constituents and a more complex structure. Different semantic theories may assign extensions to expressions and logical forms to sentences in different ways.

Many of these complexities will not concern us here. The discussion that follows should be general over many specific proposals concerning logical form, extensions of simple terms, and compositional determination. I will simply take it for granted that sentences have a logical form and contain simple terms that have an extension; that these simple terms compose complex terms, which compose the sentence; and that the extension of a complex expression (including a sentence) is at least in many cases determined by its logical form and the extensions of its parts.

To clarify terminology: On my usage, an 'expression' is any entity that has an extension (or which is a candidate for extension). For ease of discussion, I will say that when an expression is of the sort that is a candidate for an extension, but appears to lack extension, it has a null extension. A 'term' is any expression other than a sentence. Complex expressions are expressions (including sentences) that are composed of other expressions. Complex terms are terms that are composed of other terms. Simple terms are terms that are not composed of other terms.

2 Sense and extension

A simple and attractive view of meaning ties the meaning of an expression to its extension. On such a view, the meaning of a simple term is its extension, and the meaning of a complex expression is determined by the extensions of its parts. On the strongest version of this view, the meaning of a complex expression is its extension. On a slightly weaker version, the meaning of a complex expression is a complex structure involving the extensions of the simple terms that are parts of the expression. Either

way, all meaning is grounded in extension.

There are three traditional reasons for doubting this simple view of meaning. First: some simple terms (such as 'Santa Claus' and 'phlogiston') appear to lack extension. On the view above, these terms will lack meaning, or they will all have the same trivial meaning, corresponding to the null extension. But intuitively it seems clear that these terms have *some* meaning, and that their meanings differ from each others'. If so, then meaning is more than extension.

Second: in some sentences, the role of a word in determining the sentence's truth-value appears to go beyond its extension. This applies especially to sentences about beliefs and related matters. For example, it is plausible that 'John believes that Cary Grant is an actor' could be true, while 'John believes that Archie Leach is an actor' is false, even though 'Cary Grant' and 'Archie Leach' have the same extension. If so, then either the truth-value of the sentence is not determined by the meanings of the terms, or there is more to meaning than reference.

Third: there is often more than one term referring to the same thing. In such cases, the terms often seem intuitively to have different meanings. Witness 'Hesperus', the ancients' name for the evening star, and 'Phosphorus', their name for the morning star. Or take 'water' and 'H₂O', both of which refer to the same substance. If 'water' and 'H₂O' refer to the same thing, and if reference is all there is to meaning, then 'water' and 'H₂O' refer to the same thing. But intuitively, 'water' and 'H₂O' have *different* meanings. If that is so, then reference cannot be all there is to meaning.

In "Über Sinn und Bedeutung" (1891), Frege lays out the central issue roughly as follows. The sentence 'Hesperus is Hesperus' is *trivial*. It can be known a priori, or without any appeal to experience. The knowledge that Hesperus is Hesperus requires almost no cognitive work at all, and gives us no significant information about the world. But the sentence 'Hesperus is Phosphorus' is *nontrivial*. It can only be known a posteriori, by appeal to much empirical experience. The knowledge that Hesperus is Phosphorus requires much cognitive work, and gives us significant information about the world.

As Frege put it, 'Hesperus is Phosphorus' is *cognitively significant* whereas 'Hesperus is Hesperus' is not. And intuitively, this difference in cognitive significance reflects a difference in the meanings of 'Hesperus' and 'Phosphorus'. When a subject comes to know that 'Hesperus is Phosphorus', what she learns depends on what she means by 'Hesperus' and by 'Phosphorus'. And intuitively, the subject learns something different when she learns that Hesperus is Phosphorus than when she learns that Hesperus is Hesperus. If these two claims are correct, then 'Hesperus' and 'Phosphorus' have different meanings, and then meaning involves more than extension.

If meaning involves more than extension, then what is the further element? Frege held that in addition to extension (or reference), an expression also has a *sense*. 'Hesperus' and 'Phosphorus' have the same referent, but different senses. 'Water' and 'H₂O' have the same referent, but different senses. 'Cary Grant' and 'Archie Leach' have the same referent, but different senses. For all such cases, the intuitive difference in cognitive significance among pairs of terms such as these is reflected in a difference in the terms' senses.

The notion of sense has a number of important features, which I discuss in what follows. The discussion is intended as a broad and informal outline of a Fregean view, rather than as a faithful and representation of every feature of Frege's own view. Later, I will give more precise versions of some of these claims.

(1) *Every expression that has an extension has a sense.*

In "Über Sinn und Bedeutung", Frege concentrated mostly on the senses of names, holding that all names have a sense. But the same considerations apply to any expression that has an extension. Two general terms can have the same extension and different cognitive significance; two predicates can have the same extension and different cognitive significance; two sentences can have the same extension and different cognitive significance. So general terms, predicates, and sentences all have senses as well as extensions. The same goes for any expression that has an extension, or is a candidate for extension.

(2) *Sense reflects cognitive significance.*

The central feature of sense is that it is tied constitutively to cognitive significance. In the case of singular terms, Frege set out this connection as follows: two referring expressions 'a' and 'b' have different senses if and only if an identity statement 'a=b' is cognitively significant. So 'Hesperus' and 'Phosphorus' have different senses, since 'Hesperus is Phosphorus' is cognitively significant. 'Hesperus is Hesperus', by contrast, is cognitively insignificant, and two sides of the identity correspondingly have the same sense.

One can naturally generalize this test to other expressions: a pair of expressions of the same type have different senses when a statement of their coextensiveness is cognitively significant. In the case of kind terms, one can apply the same test as before: 'a' and 'b' have different senses if and only if an identity statement 'a=b' is cognitively significant. So 'water' and 'H₂O' have different senses, since 'water is H₂O' is cognitively significant. In other cases, the test will be slightly different. For general terms, one can say that 'a' and 'b' have different senses when 'All a's are b's and all b's are a's' is cognitively significant: so 'renate' and 'cordate' have different senses. For predicates, one can say that 'A' and 'B' have different senses when 'For all x, x is A iff x is B' is cognitively significant: so 'has a heart' and 'has a kidney' have different senses. For sentences, one can say that S and T have different senses when 'S iff T' is cognitively significant: so 'Hesperus is a planet' and 'Phosphorus is a planet' have different senses.

It is possible for two different expressions to have the same sense. When two words are intuitively synonymous - 'lawyer' and 'attorney'. perhaps - an identity between them is cognitively insignificant. The truth of 'lawyers are attorneys' is arguably trivial: it is knowable a priori, requires no cognitive work, and gives no significant information about the world. If so, then 'lawyer' and 'attorney' have the same sense. In a similar way, it is plausible that the sentences 'Vixens are rare' and 'Female foxes are rare' are trivially equivalent. If so, the two sentences have the same sense.

We can think of the sense of an expression as mirroring the expression's role in reason and cognition. When two expressions are trivially equivalent, they will play almost the same role in reason and

cognition, and will have the same sense. When two expressions are not trivially equivalent, they will play different roles in reason and cognition, and will have different senses. In this way, we can think of an expression's sense as capturing its cognitive significance, and as representing the "cognitive value" or "cognitive content" of the expression.

*[[Here and throughout the paper, I follow the convention of using sentence symbols such as *S*, which usually refer to a sentence, to substitute for the sentence in quotational contexts, indirect contexts, and other contexts in which it is clear that *S* is being used, not mentioned.]]

(3) *The sense of a complex expression depends on the senses of its parts.*

We saw before that the extension of a sentence (such as 'John is hot and Mary is cold') at least typically depends on the extension of the expressions it contains and on its logical form. In a similar way, the sense of a sentence at least typically depends on the senses of the expressions it contains and its logical form. The same goes for complex terms, such as 'the greatest cricket player': insofar as its extension depends on the extensions of its parts, its sense depends on the sense of its parts.

This dependence of sense on sense may proceed in much the same way as the dependence of extension on extension. We first determine the logical form of a complex expression, then determine the senses of the basic terms involved, and then compose these senses in a way that depends on the logical form, to determine the sense of the complex expression. Just how this composition works is not quite clear, but I will say more about it in what follows.

(4) *Sense determines extension.*

Frege held that the extension of a word, a complex expression, or a sentence is determined in some way by its sense. He was not entirely clear about just what sort of determination is involved here. One way of understanding the idea is the following: if two expressions have the same sense, then they have the same extension. The reverse need not be true: we have seen that it is possible two terms to have the same extension, but different sense. So there is no path from extension to sense, in general. But if this claim is right, then there is a path from sense to extension.

One question is whether sense determines extension *on its own*, or whether something else contributes to determining the extension. If the former, then it seems that the extension must somehow be present at least implicitly within the sense. But it is not easy to see how this could work, at least if sense is taken to reflect cognitive significance. The two terms 'the morning star' and 'the evening star' have the same extension, for example, but it is not clear how this sameness of extension is present within the term's senses. Similarly, a statement such as 'There are 90 chemical elements that occur in nature' may be true, but it is not clear how the truth of the sentence is determined by its sense alone. So it is natural to think that something else must contribute: namely, the world. Intuitively, the sentence above is true not just because of its sense, but because of the way the world is. And a term such as 'the morning star' refers to the planet Venus not just because of its sense, but because of the way the world is.

If this is right, the determination thesis might be put as follows: a term's extension is determined by its sense in conjunction with the world. Formulating this thesis so that it is both plausible and nontrivial is not easy: after all, is not everything determined by the way the world is? But there is at least an intuitive idea to keep in mind here, which we can return to later.

Frege also held some further theses concerning sense. These are not quite as crucial to a broadly Fregean view as the theses above, but they will be relevant to our discussion.

(5) *In indirect contexts, expressions refer to their senses.*

As we saw before, there are cases in which the truth-value of a sentence seems not to be determined by the extensions of its parts. This happens especially with sentences involving belief, and related ascriptions of attitudes. If 'John believes that Cary Grant is an actor' is true and 'John believes that Archie Leach is an actor' is false, and if 'Cary Grant' and 'Archie Leach' have the same extension, then the truth value of these sentence cannot be determined by the extensions of their parts. The same goes for many other constructions involving indirect contexts, a context where words appear inside a "that"-clause (such as 'that Cary Grant is an actor').

To deal with these cases, Frege suggested that in indirect contexts, the expressions inside the "that"-clause do not have their usual extension. They refer instead to their (usual) *senses*. So inside such a clause, 'Cary Grant' does not refer to a person, but to the (usual) sense of 'Cary Grant'. In this way, we can see that 'Cary Grant' and 'Archie Leach' have *different* extensions within these sentences, so the thesis that truth-value depends on extension is preserved.

(6) *The sense of a sentence has an absolute truth-value.*

For Frege, the sense of a sentence is a special sort of entity, a 'thought'. A Fregean "thought" is not a mental entity: it is more like what many philosophers call a *proposition*, capturing the content that a sentence expresses, when stripped of the accidental clothing of a particular language. Just as a sentence can be true or false, a proposition can be true or false. Frege held that propositions (thoughts) are the primary bearers of truth, and that sentences are true or false derivatively: a sentence is true if and only if the proposition that it expresses is true. Further, Frege held that a proposition is true or false *absolutely*. On his view, it is not possible for the same proposition to be true or false, for example at different times. If two sentences, uttered by any subjects at any times, express the same proposition, they will have the same truth-value.

This has strong consequences for the notion of sense. Many sentences can be true when uttered on one occasion, and false when uttered on another. For example, 'It is raining here now' will be false if I utter it now, but it would have been true if I had uttered it at this time yesterday. One might have been tempted to say that both of these utterances had the same sense. But if senses are propositions with absolute-truth value, they cannot. The two sentences must have different senses, and express different propositions.

(7) *The sense of an expression can vary between occasions of use and between subjects.*

It is tempting to hold that the sense of an expression is a *universal* feature of that expression: that is, that every token of an expression has the same sense. If this were right, then the sense of an expression could be seen as built into the language of which the expression is a part. But on Frege's view, sense is not universal in this fashion. One reason for this is given above: on Frege's view, the sense of a sentence such as 'It is raining here now' differs between different occasions of use. One can presumably trace this difference to differences in the sense of expressions such as 'here' and 'now' between occasions of use. Another quite different reason is tied to names. Frege says that two different users of a name such as 'Aristotle' might associate a different sense with it. He says that this should not happen in a 'perfect language', but it does happen in natural languages. For reasons like these two, Frege's view entails that one cannot attach sense to expression *types* in general. To handle cases like this, one has to attach sense to expression *tokens*.

To sum up: Frege held that expressions have senses that satisfy (1)-(7). In recent years, many philosophers have doubted this. It has been widely argued that expressions do not have Fregean senses, and there is no notion that can play the role that sense is intended to play. In particular it is widely believed that it is difficult to satisfy (2), and that it is impossible to simultaneously satisfy (1)-(4).

To flag my conclusions: I think there is a viable notion of sense such that expressions have senses that satisfy slightly modified versions of the core requirements (1)-(4). On this notion of sense, theses (5) and (6) are rejected, but (7) is accepted. I think that such a view can vindicate the spirit, if not the letter, of Frege's view.

3 What are senses?

What are senses? I have outlined a number of features that Frege attributed to senses, but this is not yet to say what a sense is. Frege's own discussion leaves the matter somewhat unclear. He says that they are not mental entities, such as the idea or image associated with an expression, for example; and he holds that they are some sort of abstract objects (inhabiting the 'third realm'). But this still leaves their nature open.

One natural suggestion is that senses are *descriptions*. Frege sometimes uses descriptions to specify senses. In talking about the sense of 'Aristotle', for example, Frege says 'It might, for instance, be taken to be the following: the pupil of Plato and teacher of Alexander the great'. And one might well hold that the sense of 'Hesperus' is something like 'the brightest object visible in the evening sky', and that the sense of 'Phosphorus' is something like 'the brightest object visible in the morning sky'.

Descriptions, on a natural understanding of the term, are linguistic entities. 'The pupil of Plato and teacher of Alexander the great' is a linguistic entity, a complex expression containing ten words. Senses obviously cannot be descriptions of this sort: descriptions have senses of their own, so we will be left either with descriptions that serve their own sense or endless chains or circles of descriptions. And more importantly, senses of this sort will not break out of the linguistic domain.

A more plausible suggestion is that the sense of an expression is the *sense* of an associated description. But even if this is true, it does nothing to tell us what senses *are*. It is also far from clear that associated descriptions of the right sort exist for all expressions. For example, 'knowledge' seems to be a paradigmatic term with a sense: some states qualify as knowledge but others do not, and one might think of the sense as encapsulating a criterion for knowledge. But there famously appears to be no description that captures such a criterion. Gettier showed that 'justified true belief' is inadequate, and all attempts at complex descriptions have failed. But even if there is no linguistic description that captures what it takes to be knowledge, this does not show that 'knowledge' has no sense.

To find a better understanding of sense, it is useful to think about the work that descriptions are doing here. The role of a description is plausibly to give us a *condition* for determining the extension of an expression, depending on how the world turns out. If the world turns out such that the brightest object in the evening sky is Venus, then 'Hesperus' will turn out to refer to Venus. If the world turns out such that the brightest object in the evening sky is Rigel Delta, then 'Hesperus' will turn out to refer to Rigel Delta. And so on. That is, the crucial property of a description is that it gives us a way of identifying an expression's extension, given full knowledge of how the world turns out.

We might then generalize by saying that a sense captures how an expression's extension depends on which world we turn out to live in. For all we know a priori, there are many ways the world could be. We might live in a world with planets visible in the sky, or we might not. We might live in a world where people play cricket, or we might not. We might live in a world where some objects travel faster than light, or we might not. We might live in a world where the liquid in the oceans is a basic atomic substance, or we might not. For all we know a priori, our world could be one of an uncountable number of different possible worlds.

This a priori ignorance about the nature of our world reflects a corresponding a priori ignorance about the extensions of our expressions. For a typical expression, such as 'Hesperus', or 'the greatest cricket player', or 'water', we cannot know what the expression refers to without much observation of the world. For all we know a priori, it might be that Hesperus is Rigel Delta, or that the greatest cricket player is Dennis Lillee, or that water is a basic element. The extensions of our expressions depend on how our world turns out. We might say: they depend on which possible world is actual.

Once we know enough about the nature of our world, we are usually in a position to know what our expressions refer to. Once we do enough astronomical work investigating the nature of the objects in the evening sky, we know that 'Hesperus' refers to Venus, not Rigel Delta. Once we know about the performances of cricket players throughout the history of the game, we know that 'the greatest cricket player' refers to Don Bradman, not Dennis Lillee. Once we know about the chemical makeup of the various substances in our environment, we know that 'water' refers to H₂O, not to a basic element. In general: once we know which possible world we live in, we can know what our expressions refer to.

We can think of this as being part of what competence with an expression involves. If a subject uses an expression, then given sufficient information about the world, the subject will be in a position to know the

extension of the expression. Furthermore, something like this will be the case *however* the world turns out: for any world, given sufficient information about that world, the subject will be in a position to determine what the extension of the expression will be *if* that world is actual. Of course in some cases the extension may be indeterminate, as it sometimes is in the actual world; but in such a case, the subject will be in a position to determine that, too.

One could put forward a thesis holding that when a subject using an expression is given sufficient information about a world, the subject is in a position to know what the extension of the expression will be if that world is actual. A full and precise version of such a thesis would require careful attention that I will not give here. But it is worthwhile to say one or two things to clarify it.

First, what counts as "sufficient information"? If we allow too much, the thesis becomes trivial: given the information that Hesperus is Phosphorus, one can trivially know that Hesperus is Phosphorus. But it is clear that in this case, no such information is required; neutrally specified information suffices. It is plausible that at least in many cases, 'qualitative' information about the distribution of physical properties, appearances, and mental states (perhaps including some indexical information) in the actual world suffices to determine an expression's extension. I will not give a precise account of the relevant information here (see Chalmers, 'Conceptual Analysis and Reductive Explanation', for a more precise hypothesis). What matters is rather the general idea that there is *some* constrained sort of information such that information of this sort can suffice to determine an expression's extension, and usually in a nontrivial way.

When is a subject "in a position to know" an expression's extension? We can say that this is the case when *sufficient reasoning* from information available to the subject will allow the subject to know the expression's extension. Here the reasoning is restricted to a priori reasoning (or to armchair reasoning, if one prefers), so further empirical observation is disallowed. And we idealize away from poor reasoning: it is not a subject's actual reasoning that matters, but rather what the subject could know given unimpaired reasoning. For example, a subject possessing the relevant information might judge that 47 plus 59 is 116, due to a miscalculation, or that Alpha Centauri is the nearest star, because they overlook the sun. But these mistakes can be corrected by better reasoning, so are no counterexample to the thesis.

If something like this is right, then a subject using an expression is in a position to know the expression's extension *given* relevant empirical information and sufficient reasoning. The second clause entails that there is a normative element here; but what matters is that the extension is within the reach of reason. This feature of language and thought is responsible for a deep link between meaning, possibility, and rationality.

4 Senses as intensions

All this suggests that an expression's sense might be seen as an *intension*: a function from possible worlds to extensions. This function takes any given possible world, and associates it with an extension within that world. The extension will be either an entity present in that world, or the null extension. We can say

that the intension is *evaluated* at a possible world, and *returns* an extension in that world.

For the Fregean, these worlds will be thought of as *epistemic* possibilities, in a broad sense: ways the world could be, for all we know a priori. And the intension of an expression can be thought of as an *epistemic* intension: it captures (very roughly) the way the extension of the expression depends on which epistemic possibility turns out to be actual. For a sentence S and a world W, for example, a useful heuristic is to ask: *if W is actual, is it the case that S?* Or to stress the epistemic nature of this conditional: *if W turns out to be actual, will it turn out that S?* If yes, then the intension of S is true at W.

Take an expression such as 'the greatest cricket player'. Let world W1 be a world where Australians never play cricket, and where Phil Tufnell scores more runs and takes more wickets than any other cricket player. Let world A be the world as it actually is, where Don Bradman's batting average is 99.94, and where the second highest batting average is around 61. For all we know a priori, world W is our world. For all we know a priori, world A is our world. If world W is our world, then the greatest cricket player is Phil Tufnell. So when evaluated at world W, the intension of 'the greatest cricket player' returns Phil Tufnell. If world A is our world (as it is!), then the greatest cricket player is Don Bradman. So when evaluated at A, the intension of 'the greatest cricket player' returns Don Bradman.

What about a term such as 'Hesperus'? Here one can tell a Fregean story in a similar way. Let world W2 be a world where the brightest object visible in the evening is Rigel Delta, and where the brightest object visible in the morning is Neptune. For all we know a priori, W2 is our world. If it turns out that W2 is our world, then it will turn out that Hesperus is Rigel Delta. So when evaluated at W2, the intension of 'Hesperus' returns Rigel Delta. If it turns out that A is our world, then it will turn out that Hesperus is Venus. So when evaluated at A, the intension of 'Hesperus' returns Venus.

The same applies to a term such as 'water'. Let world W3 be a 'Twin Earth' world, where the clear, drinkable liquid in the oceans and lakes is XYZ. For all we know a priori, W3 is our world. If it turns out that W3 is our world, then it will turn out that water is XYZ. So when evaluated at W3, the intension of 'water' returns XYZ. If it turns out that A is our world, then it will turn out that water is H₂O. So when evaluated at A, the intension of 'water' returns H₂O.

One can do the same sort of thing for a whole sentence. Just as the extension of a term depends on the way the world turns out, so does the truth-value of a sentence. So a sentence will also be associated with an intension: this time, a function from possible worlds to truth-values. This function takes a possible world, and returns a truth-value associated with that sentence in the world. This truth-value will be 'true', 'false', or perhaps neither.

Take a sentence such as 'Hesperus is Phosphorus'. One can tell a Fregean story here in a similar way. If it turns out that world W2 above (with Rigel Delta in the evening and Neptune in the morning) is actual, then it will turn out that Hesperus is not Phosphorus. So when evaluated at world W2, the intension of 'Hesperus is Phosphorus' returns 'false'. When evaluated at the actual world A, on the other hand, the intension of 'Hesperus is Phosphorus' returns true.

The intuitive characterization of epistemic intensions using the heuristics I have given here make it plausible that such entities exist, I think. Giving a truly precise definition of an epistemic intensions involves complexities that I cannot go into here. Nevertheless, it may be useful to say a few words about the foundations of the idea, and about how a more precise definition can be constructed.

The basis for epistemic intensions lies in our ability to describe and evaluate epistemic possibilities. Let us say that it is epistemically possible (in the broad sense) for a speaker that S when the speaker cannot know a priori that not- S . Then it is epistemically possible that water is not H_2O . It is also epistemically possible that our world is the XYZ-world: that is, that the clear liquid in the oceans and lakes (and so on) is XYZ. And when we reflect on the second, specific epistemic possibility, it reveals itself to us as an instance of the first epistemic possibility. That is, the epistemically possible hypothesis that the XYZ-world is actual is a specific version of the epistemically possible hypothesis that water is not H_2O .

We can see, then, that we use language to describe and evaluate epistemic possibilities in a distinctive way. Quite generally, given a specific epistemically possible hypothesis W and some more general epistemic possibility S : a speaker can say, on reflection, that the epistemic possibility that W is an instance of the epistemic possibility that S , or an instance of the epistemic possibility that not- S , or is neither. If it is the first, then the epistemic intension of S is true in W . If it is the second, then the epistemic intension of S is false in W . If it is neither, then the epistemic intension of S is indeterminate in W .

Of course any specific world must be *described* in order for a speaker to be able to evaluate it as an epistemic possibility. What the speaker really considers is whether the epistemic possibility that D is an instance of the epistemic possibility that S , where D is a description (in some sense) of W . This raises the issues of what it is to be a description of an epistemically possible world, and of whether different descriptions of the same world might give different results. I think that these questions can be answered fairly straightforwardly, and that there is a class of canonical descriptions of a world that give equivalent results, but the matter is too complex to go into here. (For this purpose, one can invoke either 'epistemically complete' or 'qualitatively complete' descriptions, as defined in 'Does Conceivability Entail Possibility',) For our purposes, the intuitive characterization of what it is for the epistemic intension of S to be true in W should suffice.

The epistemic intension of S at W corresponds to a speaker's judgment about whether the epistemic possibility that W is an instance of the epistemic possibility that S , or to the speaker's judgment about whether S is true if W is actual. But it is not the speaker's snap judgment that matters, nor any actual judgment of the speaker. Rather, it is the speaker's (potential) *rational* judgment that matters. Here we idealize away from poor reasoning, and consider judgments on ideal rational reflection.

An idealization like this can be made in a number of ways. The most obvious way is to invoke the idealization inherent in the notion of apriority. S is a priori not if a speaker knows that S , nor even if a speaker would be able to know that S on reflection, but rather if it would be possible for the speaker to know that S , using the concepts involved in S , on ideal rational reflection. So we can say that the

epistemic intension of S is true in W if a priori reasoning by the speaker could reveal the epistemic possibility of W to be an instance of the epistemic possibility that S.

All this suggests a natural definition of the epistemic intension of S for a speaker. For a given world W, let D be any canonical description of W. Then the epistemic intension of S is true at W iff the material conditional $D \rightarrow S$ is a priori; it is false at W if $D \rightarrow \sim S$ is a priori; and it is indeterminate at W if neither is a priori. Other forms of definition are also possible; for example, if someone rejects the notion of apriority, then one can appeal to a different construal of the relevant epistemic status. For this to be a truly precise definition, one needs to say more about the relevant worlds, descriptions, and epistemic status; so this is not yet a truly precise definition. (But see "The Structure of Epistemic Possibility" for a more precise treatment.)

It should be noted that to evaluate the epistemic intension of S at W, a token of S need not be present within W. None of the heuristics or definitions that I have given here rely on such a token; rather, they rely on first order claims about epistemic possibility and apriority using an expression present in the actual world. And one can often evaluate expressions at worlds with no such tokens. For example, there are epistemically possible worlds in which no-one uses language. When one considers such an epistemic possibility W, it reveals itself as an instance of the epistemic possibility that no words exist, that there are no novels, and so on. So the epistemic intension of 'words exist' is false at W, and so on. It is even arguably epistemically possible that no-one exists (as long as 'I exist' is not a priori). One can consider and evaluate that epistemic possibility in various ways; for example, the epistemic intension of 'Someone exists' will be false there. Of course some sentences may yield indeterminate results at these epistemic possibilities, but all that matters is that some sentences yield determinate results there.

All of this arises from the fact that when a speaker is given the right sort of information about the actual world and engages in the right sort of reflection, conclusions about extension are within reach of reason. And when a speaker is given the right sort of information about an epistemic possibility, and engages in appropriate reflection, conclusions about extension within that epistemic possibility are within reach of reason.

We can use this to articulate, in this framework, the Fregean notion of 'grasping' a sense. We can say that a subject grasps an intension when the subject is in a position to *evaluate* that intension: that is, when sufficient reasoning will allow the subject to determine the value of the intension at any world. Again, this does not require that the subject will actually determine the correct extension when a relevant epistemic possibility is specified, but it does require that the extension is within the grasp of reason. If what I have said here is right, then whenever a subject uses an expression competently, the subject will grasp the expression's epistemic intension.

Occasionally, the epistemic intension of a term will be the same as that of an associated description. For example, the epistemic intension of 'Neptune' as used by Leverrier was arguably the same as that of 'the object causing the perturbation of the orbit of Uranus'; and the epistemic intension of 'bachelor' is arguably the same as that of 'unmarried man'. In these cases, there is a sense in which the epistemic intension can be "captured" by a description.

This does not hold in general. For many or most terms, there may be no description (and certainly no short description) with the same epistemic intension as the term. We saw this in the case of 'knowledge', and the same applies to most names. In these cases, the best one can hope for is a description whose epistemic intension approximates that of the original term: as with 'justified true belief' for 'knowledge', or 'the clear drinkable liquid in the oceans and lakes' for water, and so on. These descriptions may give one a rough and ready sense of how a term's epistemic intension functions, but they do no more than that. Usually there will be at least a small subset of epistemic possibilities (such as the Gettier cases, in the case of 'knowledge') in which the epistemic intension of the original term and of the description come apart.

So in general, there is no reason to think that an epistemic intension can be captured by a description. And more generally, there is no reason to think that grasping an epistemic intension requires any sort of descriptive articulation of a concept by a subject. The epistemic intension is a function, not a description. It is revealed in a subject's rational evaluation of specific epistemic possibilities, not in any sort of explicit definition. Even where such a definition exists, a subject need not be able to articulate it to grasp the epistemic intension. Indeed, we usually evaluate the plausibility of such definitions precisely by deploying our prior grasp of a term's epistemic intension, to see how whether the definition gives the right results in specific cases. (Witness the literature on the definition of 'knowledge'). So epistemic intensions are more basic than descriptions, and should not be assimilated to them.

Epistemic intensions can be associated with any expression that is a candidate for extension. Given the general type of the expression (singular term, general term, etc), it will be constrained to have a certain sort of extension (individual, class, etc). The intension of such an expression will be a function from worlds to the appropriate sort of extension. So the intension of a singular term is a function from worlds to individuals; the intension of a general term is a function from worlds to classes; the intension of a sentence is a function from worlds to truth-values; and so on.

This thesis can be considered a basic postulate of the Fregean framework:

(1) Every expression has an intension, which returns at a world an extension of the type appropriate for the expression.

We can add to this two further postulates:

(3) If the extension of a complex expression E depends on the extensions of its parts, then at a world W, E's intension returns an extension that depends in the same way on the extensions returned by the intensions of its parts.

(4) At the actual world, the intension of E returns the (actual) extension of E.

The Fregean could simply take these theses as basic postulates. But it is also possible to see how they fit

with our intuitive understanding of an epistemic intension. For (2): part of that understanding is that any expression can be evaluated at arbitrary epistemic possibility, yielding the same sort of extension that it yields in the actual world. For (3): the dependence of the extension of a whole on the extension of parts is presumably not an accident about the actual world, so will hold across arbitrary epistemic possibilities. For (4): given appropriate information about the actual world, a subject will be in a position to determine the expression's actual extension.

Given (1), (3), and (4), and the equation of senses with intensions, it is easy to see how the Fregean requirements (1), (3), and (4) follow. Requirement (1), that every expression has a sense, follows immediately from (1). Requirement (3), that the sense of a sentence depends on the sense of its parts, follows from (3) along with the thesis that the extension of a sentence depends on the extension of its parts. Note that even if there are some cases where the latter thesis fails (such as belief contexts), requirement (3) can still be satisfied as long as the extension of a sentence depends on the extensions *and the intensions* of its parts. For then it is plausible that the extension of a sentence at a world will depend on the extension of its parts at a world along with the intensions of the parts, so the intension of the sentence will depend on the intensions of the parts.

As for requirement (4), that sense determines extension: a version of this immediately follows from (4). The relevant version is the second version given earlier, holding that sense determines extension in combination with the world. To determine an expression's extension, one simply evaluates the expression's intension at the actual world. One can naturally think of the intension as supplying *criteria* for determination of extension: in combination with the actual world, these criteria will determine an extension.

What about the crucial requirement (2), that sense reflects cognitive significance? Here matters are complicated by the fact that we lack a precise definition of 'cognitive significance'. There is a natural understanding of cognitive significance that fits well with the intensional framework, however. We can say that a sentence *S* is cognitively insignificant when *S* is knowable a priori, or with justification independent of experience. And *S* is cognitively significant when it is not knowable a priori. 'Hesperus is Hesperus' is knowable a priori, so it is cognitively insignificant. 'Hesperus is Phosphorus' is not knowable a priori, so it is cognitively significant.

This understanding of cognitive significance is not quite Frege's. On Frege's account, a priori knowledge can be cognitively significant: the knowledge that $59+46$ is 115 is cognitively significant, for example, because this knowledge requires some cognitive work. It is very hard to articulate this notion precisely, however, and it is not clear that there is a useful precise notion nearby. In any case, the definition in terms of apriority is at least not too far from Frege's central notion, and it handles most of Frege's central cases, which involve a posteriori knowledge. So it is this understanding that I will use.

For senses to reflect cognitive significance in this sense, it was required (among other things) that an identity ' $a=b$ ' is cognitively significant if and only if ' a ' and ' b ' have different senses, and that two sentences *S* and *T* have the same sense if and only if the material conditionals 'If *S* then *T*' and 'If *T* then *S*' are cognitively insignificant. Recasting these claims in terms of intensions and apriority, we require the

following: An identity ' $a=b$ ' is a priori if and only if ' a ' and ' b ' have the same intension; and two sentences S and T have the same intension if and only if the material biconditional ' S iff T ' is a priori.

Both of these claims are entailed by the following principle:

(2) S is a priori if and only if its intension is true at all worlds.

>From (2), it follows that ' $a=b$ ' is a priori iff its intension is true at all worlds. By (3) and the extensional semantics of identity, the intension of ' $a=b$ ' is true at a world iff the intensions of ' a ' and ' b ' return the same extension there. So ' $a=b$ ' is a priori iff the intensions of ' a ' and ' b ' return the same extensions at all worlds; that is, iff ' a ' and ' b ' have the same intensions.

>From (2), it also follows that ' S iff T ' is a priori iff its intension is true at all worlds. By (3) and the semantics of material conditionals, the intensions of ' S iff T ' is true at a world iff the intensions of S and T return the same truth-value there. So ' S iff T ' is a priori iff the intensions of S and T return the same truth-values at all worlds; that is, iff S and T have the same intensions.

So from (1)-(4), versions of the crucial Fregean requirements (1)-(4) can be satisfied. The question then becomes: why believe that (1)-(4) can be satisfied? The crucial claim is clearly (1). So, why believe that expressions can be associated with intensions that satisfy (1)?

The central reason stems from the Fregean understanding of these intensions and of the associated worlds. As we saw above, the worlds represent *epistemic* possibilities: ways the world might be, for all we know a priori. And the intensions are *epistemic* intensions: they capture a subject's idealized judgments about how an expression applies to an epistemic possibility, under the hypothesis that it is actual. On this understanding, a strong connection between intensions and epistemic notions such as apriority is built into the framework.

So consider the left-to-right direction of (1): if a statement is a priori, its epistemic intension is true at all worlds. This direction is straightforwardly plausible. If S is a priori, then for *any* epistemic possibility W : if W turns out to be actual, it will turn out that S . We might say that for all W , the epistemic possibility that W is actual is an instance of the epistemic possibility that S . So the epistemic intension of S is true in all worlds.

This is brought out by cases such as ' $\text{Hesperus is Hesperus (if it exists)}$ ', ' $\text{All bachelor are unmarried men}$ ', and even ' $\text{The meter stick in Paris is one meter long (if it exists)}$ ', assuming that these are a priori. In each of these cases, there is no epistemic possibility of falsehood. No matter how the world turns out, it cannot turn out that Hesperus is not Hesperus, or that the meter stick in Paris is longer or shorter than one meter. So in every epistemic possibility W , the epistemic intensions of these sentences are true.

Something similar applies in the right-to-left direction, which we can consider in the contrapositive form: if S is not a priori, then there is some world in which its epistemic intension is not true. This fits the

familiar cases. 'Hesperus is Phosphorus' is a posteriori, and its epistemic intension, as we saw above, is false in a world W2 where the objects visible in the morning and evening sky differ. 'Water is H₂O' is a posteriori, and its epistemic intension is false in the 'Twin Earth' world W3. And this is no accident: the aposteriority of these statements seems to be reflected in the existence of these worlds in which the epistemic intension is false.

If S is not a priori, it is clearly epistemically possible that not-S. So all we need for the right-to-left direction of (1) is the following claim: if it is epistemically possible that T, then there is an epistemically possible world W in which the epistemic intension of T is true. This claim is suggested very strongly by the above examples, but it is not completely trivial.

Someone might hold in response that there are counterexamples to the claim. They might hold that 'God exists' (S1) is necessary but not a priori, for example; or they might hold that the Continuum Hypothesis (S2) is necessary but not a priori. And they might suggest that in these cases (unlike cases involving water, Hesperus, and so on) there is no world W such that if W is actual, S1 is false or S2 is false.

These claimed counterexamples would be highly controversial, and a Fregean could simply deny them as contrary to his or her principles, holding that the opponent is assuming a false view of the cases, and perhaps tacitly assuming a false view of some of the crucial notions: apriority, possibility, intensions. But it is useful to probe just why a Fregean would deny them. The Fregean could deny them in at least two ways, depending on the relevant conception of possibility.

There are two ways in which the Fregean might understand the worlds that are considered as epistemic possibilities and over which epistemic intensions are defined. First, they might be a space of worlds that is *independently characterized*, perhaps as the space of metaphysically possible worlds. In this case, it becomes a substantive thesis that when S is epistemically possible, there is a world in which the epistemic intension of S is false. The thesis is plausible for familiar cases (including 'water' and 'Hesperus' cases), but an opponent might hold that in some cases, it is false. This might happen for reasons such as those in the case above, where an opponent could hold that there are not enough metaphysically possible worlds to go around.

On this understanding, for the Fregean to hold onto principles such as (1), she will have to deny the opponent's analysis of the cases. She might hold that God's existence cannot be necessary, for example, or perhaps that it is a priori; and she might hold that the Continuum Hypothesis is either a priori or indeterminate. This will require substantive argument. And the postulated connection between the independently characterized space of worlds and epistemic possibility will be a substantive thesis; it will have implications for just what is metaphysically possible, for example. Of course the thesis may well be plausible for all this. A Fregean might even argue for it directly, perhaps by noting that there are no clear counterexamples, and by arguing that there are constitutive connections between notions of metaphysical and epistemic possibility.

On the second understanding, the worlds involved are not independently characterized, but are understood as epistemic possibilities from the start. The Fregean might postulate or construct a space of

worlds understood as 'maximal epistemic possibilities', for example. These might intuitively correspond to maximally specific epistemically possible hypotheses about the way things are, hypotheses from which a priori reasoning can settle everything there is to be settled. This space might be defined quite independently of notions of 'metaphysical possibility'.

On this understand, a Fregean is free to accept much of the opponent's analysis of the cases above. She is free to accept that 'God exists' is a posteriori but metaphysically necessary, for example. If she accepts this, she will simply insist that there is an epistemically possible world in which the epistemic intension of 'God exists' is false. This will follow naturally from the definition of those worlds. Once epistemic possibility is separated from metaphysical possibility, the opponent will have no obvious reason to deny this. On this understanding, the truth of (1) does not involve commitment to the substantive theses about metaphysical possibility above, and it will be well-protected from counterexamples.

(A third possibility is that the Fregean could accept both the opponent's analysis of the cases and an independently characterized notion of possibility, and accept that (1) is not true across the board. It might be held that there are certain special domains where it fails, but that it holds in the most domains, or in certain constrained domains. This is not out of the question, but I will set it aside here.)

My view is that the Fregean can go in either of these two ways. I think that (1) is plausible even given an independently constrained notion of metaphysical possibility, and that it can be argued for directly; but this is a nontrivial matter. If someone doubts this, however, the second option is available. This has the advantage of making fewer substantive commitments. Here, the main burden is making the case for the relevant space of epistemically possible worlds. This is also nontrivial, but I think that it can be done in a reasonably straightforward way. (See my 'Materialism and the Metaphysics of Modality' for arguments that cash out the first strategy, and 'The Structure of Epistemic Possibility' for a construction that cashes out the second.)

In any case, we can see that the Fregean is not being unreasonable in accepting (1). It appears to fit the clear cases, and there are natural ways to respond to one who doubts it. So it seems that the claim that there are intensions of some sort that satisfy (1)-(4) is well-motivated. If there are such intensions, then they can function as senses that satisfy (1)-(4).

To clarify what I have and have not done: I have not here tried to precisely define epistemic intensions, in the sense of giving a precise recipe for evaluating a sentence S's intension at a world W. I have also not tried to give a knockdown argument that (1)-(4) can be satisfied. Rather, I have simply outlined a certain sort of Fregean view, and I have tried to make it plausible that intensions satisfying (1)-(4) exist.

If this is correct, a Fregean can hold that there is at least a *prima facie* case for a notion of sense that satisfies (1)-(4). But of course there have been numerous arguments *against* Fregean sense in recent years. So we now need to examine these arguments to see whether they have any force against the conception just outlined. I will focus on four main arguments: what we might call the argument from indexicality, the modal argument, the epistemological argument, and the argument from variability.

5 The argument from indexicality

The first objection is not so much an explicit argument as a set of considerations put forward by a number of philosophers, especially John Perry (1977). These considerations revolve around Frege's treatment of indexicals, such as 'I', 'here', and 'now'.

Recall that Frege held that the sense of a sentence has an absolute truth-value. This entails that if two utterances of a sentence express the same sense, they must have the same truth-value. But it is clear that certain indexical sentences, such as 'It is now Saturday' can be uttered truly at one time and falsely at another time. So on Frege's picture, these two sentences must have different senses.

If the sense of the sentence depends on the sense of its parts, then some part of the sentence must have a sense that differs between the two occasions of utterance. The obvious source of the difference is the indexical expression 'now'. So Frege's view entails that such an indexical expression has different senses on different occasions of utterance. The same goes for many other indexicals: 'I', 'here'. 'today', 'tomorrow', and so on.

It is very hard to see how this is supposed to work. One idea is that the sense of such an expression should build in its referent. If so, my utterance of 'now' today has a sense that builds in a particular Saturday, and my utterance of 'now' tomorrow has a sense that builds in a particular Sunday. Similarly, my utterances of 'I' have a sense that builds in a particular individual, David Chalmers, while your utterances have a sense that builds in a different individual.

It seems that something like this is required to preserve Frege's claims, but this also seems quite contrary to the overall spirit of Frege's framework. Sense is usually quite distinct from reference, and it is not clear why this case should be different. When I use an expression such as 'now' or 'today', the referent does not seem to be reflected in the cognitive significance of the expression for me. I might have no idea what day today is; and the day might change without it making any special difference to my cognition. So there seems to be a strong tension between this sort of claim and the general spirit of (2).

How will the intensional framework deal with these matters? Let us consider a sentence such as 'It is raining here now'. If such a sentence determines a function from worlds to referents, there is an immediate problem. I can utter the sentence truly today, and falsely yesterday. But both days, I inhabit the same world: at least, the same 'objective' world A. So if the intension of a sentence is a function from (objective) worlds to truth-values, then as before, the two utterances must have different intensions. And as before, the two tokens of 'now' must have different intensions.

It is once again quite unclear how this should work. One might suggest that 'now' should pick out a particular time (time t , the specific time of utterance) in all worlds. But then a sentence such as 'It is now time t ' will have a necessary epistemic intension. But this sentence is clearly a posteriori (and cognitively significant), so it must have a contingent epistemic intension. So this will not do. And it seems that any other intension will have a similar problem.

I think there is only one natural way for the intensional framework to deal with this issue. A proponent of this framework should deny that intensions are functions from objective worlds to extensions. Rather, intensions must be seen as functions from *centered* worlds to extensions. Here, a centered world is a world marked with a 'center'. where the center consists of an individual and a time present in that world.

This idea can be motivated in a natural way. On the Fregean intensional framework, worlds are supposed to represent a sort of maximal epistemic possibility. But it is a familiar idea (this time from the work of Perry (1979), among others) that an objective description of the world leaves some matters epistemically open. When I lie awake in the middle of the night, then even if I had a full objective description of the world, I might still wonder 'what time is it now?', and I might not be able to settle this matter from the information available. Or I might have a full objective description of the world, but not know which individual in that world is *me*. So an objective description of the world is not an epistemically complete description of the world. To make it epistemically complete, the description also needs *locating information*: a 'you are here' marker, indicating which individual is *me* and what time is *now*. This sort of epistemic possibility is best represented by a centered world.

Once epistemic possibilities are represented as centered worlds, we can deal with the problem straightforwardly. We can say that the epistemic intension of 'I' picks out the individual marked at the center of any given world, and that the epistemic intension of 'now' picks out the time marked at the center. The epistemic intension of 'today' will pick out the day containing the time marked at the center of any given world, and the epistemic intension of 'tomorrow' will pick out the following day. These intensions will be common to all occasions of use of these expressions.

When a subject uses an expression, the actual extension of the expression will be given by evaluating the expression's intension at the centered world inhabited by that subject: a world corresponding to the actual world centered on that subject and on the time of use. When two different subjects use an indexical expression such as 'I'. they will inhabit two different centered worlds: one centered on the first subject, and one on the second. So the epistemic intension of 'I' will pick out different actual extensions for each. For each subject, the intension will pick out himself or herself. Something similar applies to 'now': when this expression is used at different times, the intension will be evaluated at different centered worlds, and will always return the time of use.

Something like this can also help with terms that are not obviously indexical, such as 'water'. If I am given a full objective specification of an epistemic possibility, and am told that it contains a planet where the watery stuff is H₂O and a planet where the watery stuff is XYZ, then I may not be in a position to know what the extension of 'water' is. To know that, I need to know which planet I am on. But if I am also given locating information, in which the center of my world is marked (e.g. on the planet with H₂O), then there is no problem. I am now in a position to know which environment is *my* environment, to know which substances *I* am causally related to, and so on. So as long as I can derive the relevant objective information (about appearance, behavior, distribution of various substances), I will have no problem determining that if this centered world is actual, then water is H₂O. So the epistemic intension of 'water' will return H₂O at this world.

There are a couple of subtleties to the use of centered worlds. One is the following: it is arguably not a priori that I exist (I know this through experience), but the epistemic intension of 'I exist' is true in all centered worlds as defined. To deal with this, one should make the marking of a subject and time *optional*: some centered worlds have no marked subject and time, or perhaps mark just one but not both. In a centered world without a marked subject, the epistemic intension of 'I exist' will be false. The other side of this coin is that there may be expressions that require the marking of further entities at the center of a world: some demonstratives ('that object') may require marked experiences, for example, and some special cases ('this thought') might require marked thoughts. These cases are not crucial to the current discussion, however, so we can mostly stay with worlds in which at most a center and a time is marked.

The introduction of centered worlds to the Fregean framework has one major consequence. The sense of a sentence will no longer have an absolute truth-value. When I say 'It is raining here now' yesterday and today, my utterance has the same epistemic intension both times, but it is false yesterday and true today. So the intension is not true absolutely, or false absolutely. It is true or false only relative to a subject and a time. So the Fregean requirement (6) fails.

It is not clear what Frege would have thought of this. The requirement that the sense of a sentence (a 'thought') be an absolute bearer of truth was very important to him. Still, it is widely held that Frege's treatment of indexicals needs major repair. And it seems to me that giving up thesis (6) does not do any significant damage to a broadly Fregean framework, and it allows one to preserve the crucial connection with cognitive significance and the determination of reference. So it may be that giving up this thesis is the best way to preserve a framework that is in the broad spirit of Frege's.

This adjustment also entails that we must give up on Frege's thesis (5), that expressions in belief contexts refer to their senses. Consider a belief attribution such as 'John believes that I am British'. Here the sense of 'I' is the epistemic intension that picks out the individual at the center of a given world, and the sense of 'I am British' is true only in worlds where the individual at the center has a certain national origin. If John were to entertain a belief with that sense, then he would attribute that national origin to *himself*. But it is clear that this is not what John does when he believes that I am British. So on the current understanding of sense, thesis (5) entails that the wrong sort of belief is attributed to John. So we must give up on thesis (5).

I do not think that this is a high cost to pay. It is widely held that thesis (5) must be rejected, for a number of different reasons. The thesis yields an attractively elegant analysis of belief attributions, but on a close analysis, it seems that belief attributions are more subtle than the thesis suggests. The analysis of belief attributions was only a subsidiary element of Frege's view, however, and it is clear that giving it up preserves the broad spirit of his view. Of course we still need a good analysis of belief attribution, but that is a subject that needs much discussion in its own right. (I think that epistemic intensions can still be exploited, but in a more indirect way. See "The Components of Content".)

6 The modal argument

Perhaps the best-known argument against a Fregean view of language is the modal argument of Kripke (1972). Kripke's argument is concerned mostly with names, and also with natural-kind terms. He is arguing against 'descriptive' views of reference in the first instance, on which names are semantically akin to descriptions, but the arguments are generally taken to have force against any Fregean view.

Kripke argues that names cannot be equivalent to descriptions as follows. Take any name, such as 'Aristotle', and any description, such as 'the last great philosopher of antiquity'. Then it might have been that Aristotle was not the last great philosopher of antiquity; he might have died while an infant, for example. Something similar applies to any other description D that seems a likely candidate to capture the sense of the name: for any such description D, we judge that it might have been that Aristotle was not D. So 'Aristotle' is not semantically equivalent to the description D.

Something similar applies in the case of 'Hesperus'. A Fregean might hold that 'Hesperus' is semantically equivalent to 'the brightest object visible in the evening sky', or some such. But Kripke argues that it might have been that Hesperus was not the brightest object visible in the evening; it might have been destroyed millenia ago, or it might have been struck by a comet and left the solar system. So again, it seems that a name is not semantically equivalent to a description.

Kripke uses this to argue that it is not *necessary* that Hesperus is the brightest object visible in the evening sky, since it might have been otherwise. Or as it is sometimes put: it is not *metaphysically necessary* that Hesperus is visible in the evening, and it is *metaphysically possible* that it is not. In a similar way, he argues that there are *possible worlds* in which the evening star is not Hesperus: in a world where Venus was knocked off course by a comet and in which another object is visible in the evening, the evening star is not Hesperus but some other object.

He argues in a similar way that there are no possible worlds in which Hesperus is not Phosphorus. If Venus had been visible only in the morning with something else visible in the evening, this would have been a scenario in which Hesperus was not visible in the evening, and not a scenario in which Hesperus was not Phosphorus. He argues that Hesperus and Phosphorus are the same object (the planet Venus) in all worlds in which they exist. So 'Hesperus is Phosphorus (if they exist)' is necessary.

Kripke puts this by suggesting that names are *rigid designators*, picking out the same object in all possible worlds. Most descriptions, in contrast, are not rigid: they pick out different objects in different possible worlds. So names are quite unlike descriptions. Where names are concerned, Kripke's view is closer to the simple view than to the Fregean view: the meaning of a name is simply its referent, and not any associated description or sense.

Kripke gives related arguments concerning natural kind terms. Take the case of 'water' (Kripke does not concentrate on this case, but it is useful to stick to the familiar case). It can be argued that water might have behaved and appeared quite differently from the way in which it actually behaves and appears: it might never have appeared in liquid form, for example (witness the possibility of ice in the actual world). More generally, for any description D of water's macroscopic properties, it can be argued that if H₂O had

not been D, water would not have been D. So it is not necessary that water is D, and it seems that natural kind terms are not equivalent to descriptions.

In a similar way, it can be argued that something might have satisfied any such description D without being water. In Putnam's Twin Earth world, a different chemical substance XYZ has all the superficial properties of water. But Putnam argues that this substance is not water. That is, in a counterfactual scenario in which XYZ was watery, XYZ would not be water. Rather, water is necessarily H₂O. If so, 'water' is akin to a rigid designator: it picks out H₂O in all worlds.

Do Kripke's arguments have any force against the intensional framework I have outlined here? Do they show, for example, that names or natural kind terms do not have epistemic intensions that satisfy (1)-(4)? One might think that they do. I have argued that Hesperus has an epistemic intension that picks out something like an object visible in the evening, in any given world. So in a world where Mars rather than Venus is prominent in the evening sky, the epistemic intension will pick out Mars. But Kripke argues that Hesperus picks out Venus in all worlds, and that it needs not pick out the evening star, or anything like it.

The conflict is only superficial, however. Kripke takes care to distinguish metaphysical possibility from epistemic possibility. And he allows that it is epistemically possible that Hesperus is not Venus; he simply denies that it is metaphysically possible. Kripke allows that it might *turn out* that Hesperus is not Phosphorus; and he can allow that if it turns out that Venus was never visible in the evening but that Mars was, then it may turn out that Hesperus is not Phosphorus but Mars. So his argument is entirely compatible with 'Hesperus' having an *epistemic* intension that functions roughly as I have suggested. Much the same applies to 'water', and to other relevant terms.

This response is not ad hoc. The sort of possibility that is most relevant to a Fregean view is clearly epistemic possibility. When one thinks about sense in intensional terms, one thinks of it as giving criteria for the extension of an expression *depending on how the world turns out*. If the world turns out one way, it will turn out that water is H₂O; if the world turns out another way, it will turn out that water is XYZ. Nothing in Kripke's modal argument gives any reason to deny this. And the notion of sense was always tied to epistemic notions such as apriority, not to notions such as 'metaphysical necessity'. These connections are entirely preserved, even in the light of Kripke's argument.

The other crucial property of sense was that it determines an expression's extension in the actual world. Again, Kripke's argument does nothing to suggest that this is not the case. The epistemic intension determines an expression's actual extension when evaluated at the actual world. For all Kripke has said, it also determines an expression's extension under all epistemically possible hypotheses about the actual world.

What an epistemic intension does not do, if Kripke's arguments are correct, is determine an expression's extension when evaluated in explicitly counterfactual scenarios. When we consider these scenarios, we are not considering them as epistemic possibilities: as ways things might be. Rather, we are acknowledging that the character of the actual world is fixed, and are considering these possibilities in the subjunctive mood: as ways things might have been. That is, rather than considering the possibilities *as*

actual (as with epistemic possibilities), we are considering them *as counterfactual*. If Kripke is right, then evaluation in this sort of explicitly counterfactual context works quite differently from the evaluation of epistemic possibilities. This point still needs explaining.

It is striking that all of Kripke's conclusions concerning modality are grounded in claims concerning what might have been the case, or what could have been the case, or what would have been the case had something else been the case. Kripke is explicit (1980, pp. 36-37) in tying his notion of necessity to these formulations, and almost all of his arguments for modal claims proceed via these claims. What all these formulations have in common is that they involve scenarios that are acknowledged not to be actual, and that are explicitly considered as counterfactual scenarios.

All these claims are *subjunctive* claims, not in the syntactic sense, but in the semantic sense: they involve hypothetical situations that are considered as counterfactual. The paradigm of such a claim is a subjunctive conditional: 'if P had been the case, Q would have been the case'. We can say that all these claims involve a *subjunctive context*, where a subjunctive context is one that invokes counterfactual consideration. Such contexts include those created by "might have", "would have", "could have", or "should have" (on the non-epistemic readings of these phrases), subjunctive conditionals involving "if/were/would be" or "if/had/would have", and other phrases. In Kripke's sense of "possible" and "necessary", where it taken to be possible that P if it might have been the case that P, then modal contexts such as "It is possible that" and the like can be considered as subjunctive contexts, too.

Kripke's central point against the description theory was that names and descriptions function differently in modal contexts: for a name 'N' and a description 'D', it is necessary that D is D (if it exists), but it is not necessary that N is D. We can put this somewhat more precisely by saying that names and descriptions function differently in subjunctive contexts. And more generally, names and descriptions seem to behave differently under subjunctive evaluation of hypothetical possibilities. How can a Fregean view handle this phenomenon?

I have not argued that names are equivalent to descriptions. I do allow, however that a name and a description can have the same sense, at least approximately and in some cases, as the case of 'Hesperus' suggests. If so, then if Kripke's point is accepted, we must explain the different truth-value of subjunctive sentences by appealing to something other than a difference in sense. So the question is: why is 'it might have been that Hesperus was not the evening star' true, while 'it might have been that the evening star was not the evening star' is false, given that 'Hesperus' and 'the evening star' have (roughly) the same sense?

There are a number of ways in which one might try to explain this. First, one might appeal to a difference in underlying logical form between subjunctive sentences containing names and descriptions. One could hold that in subjunctive contexts, names always take wide scope, so that the sentence involving 'Hesperus' above has the logical form of 'Hesperus is such that it might not have been the evening star'. Or one could hold that names always involve an unarticulated 'actual'. so that the sentence has the logical form of 'it might have been that the actual Hesperus was not the evening star'. Once this logical form is in place, then substituting 'the evening star' for 'Hesperus' yields the same truth-value. On such a view, the

difference in the original sentences was solely due to a difference in underlying logical form, not in semantic content.

These explanations merit considerable discussion, but I will set them aside here. The second sort of explanation appeals to a semantic difference between names and descriptions, in some aspect that goes beyond their sense, and that affects how they function in subjunctive contexts. On such a view, a name may have sense and extension, but sense and extension do not *exhaust* the meaning of a name.

The simplest addition to the Fregean framework would be a semantic feature that is part of the meaning of all names but not part of the meaning of descriptions. When present in the meaning of an expression, this feature would indicate that within counterfactual contexts, the expression contributes its *actual* extension. There will be a corresponding semantics of counterfactual contexts, such that the presence or absence of the feature is relevant to the truth-value of counterfactual statements. (Recanati (1993) proposes a feature he calls 'REF' that works something like this, though he conceives of it differently.) This will explain the difference between names and descriptions.

I favor a more general semantic explanation. On this account, every expression is associated with *two* intensions: one governing its application to epistemic possibilities, and one governing its application to explicitly counterfactual possibilities. The first of these is the epistemic intension. The second is a *subjunctive intension*. Like the epistemic intension, this is a function from worlds to extensions. But here the worlds in question are seen as counterfactual metaphysical possibilities, and expressions are evaluated in these worlds in the way in which we evaluate counterfactual scenarios.

Just as the epistemic intension mirrors the way that we describe and evaluate epistemic possibilities, the subjunctive intension captures the way that we describe and evaluate subjunctive counterfactual possibilities. To evaluate the subjunctive intension of a sentence *S* in a world *W*, one can ask questions such as: if *W* were the case, would *S* be the case? For example, if *W* is a world in which Venus was knocked off course by a comet and in which Mars was prominent in the evening, we can say the following. If *W* were the case, Hesperus would not be visible in the evening; if *W* were the case, Hesperus would still be Phosphorus, and would still be Venus; if *W* were the case, Hesperus would not be Mars. So the subjunctive intension of 'Hesperus is Phosphorus' is true in *W*.

The same sort of thing applies to a natural kind term, such as 'water'. If *W* is the Twin Earth world, we can say: if *W* were the case, water would still be H₂O, not XYZ. So the subjunctive intension of 'water is XYZ' is false in that world, and the subjunctive intension of 'water' picks out H₂O in that world. One can tell a related story for a general term such as 'cat', whose subjunctive intension will pick out a class of members of a particular biological species in all worlds, and for a property term such as 'hot', which will arguably pick out a certain sort of molecular motion in all worlds. (Note that at least in the 'cat' case, the extension is a different class in each world, and is not always the actual extension. This can be handled straightforwardly by a subjunctive intension, but it is harder for the other accounts above to handle, as they rely wholly on projecting the actual extension.)

Something similar applies to an indexical such as 'I'. The epistemic intension picks out the being at the center of a world, but the subjunctive intension does not. The subjunctive intension of my use of 'I' picks out me (David Chalmers) in all worlds. I can say 'if David Chalmers were in a coma, then I would be in a coma', and so on. In evaluating counterfactual scenarios, 'I' always picks out David Chalmers (though of course, he need not be *called* that). So we do not need a center to evaluate the subjunctive intension of 'I'; and more generally, the worlds in the domain of subjunctive intensions can be taken to be standard uncentered worlds, not centered worlds.

Just as every expression has an epistemic intension, every expression will have a subjunctive intension. This intension will be a function from worlds to extensions of the appropriate sort: individuals, truth-values, and so on. And as long as the extension of a sentence depends on the extensions of its parts, the subjunctive intension of a sentence will depend on the subjunctive intension of its parts.

A name and a description may have similar epistemic intensions and similar extension, but their subjunctive intensions will be very different. This can be seen by examining world W above. Here we can say that if W were the case, then Hesperus would still be Venus, but the evening star would be Mars. So the subjunctive intensions of 'Hesperus' and 'the evening star' are distinct. The former picks out Venus in all worlds, while the latter picks out something that is visible in the evening in all worlds. More generally, the subjunctive intension of a name picks out its actual extension in all worlds, while the subjunctive intension of a description picks out the object that satisfies the description in all worlds.

The subjunctive intension of a name depends directly on its actual extension. The extension of a name can usually not be known a priori, so the subjunctive intension cannot be known a priori, either. If we lack relevant empirical information about the actual world, we might be unable to evaluate an expression's subjunctive intension at a counterfactual world, even given a detailed specification of that world. For similar reasons, two names (such as 'Hesperus' and 'Phosphorus') may have the same subjunctive intension without the subject knowing that a priori. The same goes for natural kind terms and indexicals. So subjunctive intensions do not reflect the cognitive significance of the expressions involved. This contrasts with the subjunctive intensions of descriptions, which can often be evaluated a priori (at least if the descriptions contain no names, natural kind terms, or indexicals), and it contrasts with epistemic intensions, which can generally be evaluated a priori for any expression.

Subjunctive intensions are most directly relevant to the evaluation of subjunctive sentences. When an expression occurs in a subjunctive context ('it might have been that S'; 'if S had been the case, T would have been the case'; 'if A were B, C would have been D'; and so on), its subjunctive intension is used in evaluating the truth-value of the sentence. For example, 'it might have been that S' (at least in the relevant sense of 'might have been') will be true iff the subjunctive intension of S is true at some world. 'If S had been the case, T would have been the case' is true roughly iff the nearest world that satisfies S's subjunctive intension satisfies T's subjunctive intension. And so on.

Subjunctive intensions are also relevant to the evaluation of modal contexts, at least when these are interpreted the Kripkean way. 'It is necessary that S' will be true iff S's subjunctive intension is true at all worlds. 'Possibly, T' will be true if T's subjunctive intension is true at some world. This is what we would

expect, given that 'It is possible that T' will be true precisely when 'It might have been that T' is true.

By analogy, epistemic intensions are relevant to the analysis of analogous epistemic contexts. The indicative conditionals 'If P is the case, then Q is the case' or 'If it turns out that P, it will turn out that Q' are true (or assertible) for a subject approximately when the epistemically nearest world that satisfies P's epistemic intension also satisfies Q's epistemic intension, where "epistemically nearest" is defined in a way that depends on the subject's knowledge or perhaps beliefs. (See "The Tyranny of the Subjunctive".) And 'it is a priori that P' will be true if P has a necessary epistemic intension.

(The material from here on is a somewhat more elaborate development of this framework, and can be skipped by those who prefer to move on.)

The subjunctive intension of an expression is always determined by the epistemic intension of the expression and by the character of the actual world. Because of this, it is possible in principle to associate a *two-dimensional intension* with an expression, which captures how its subjunctive intension will vary, depending on which epistemic possibility turns out to be actual. This two-dimensional intension can be thought of as a function from pairs (V, W) of epistemic possibilities and metaphysical possibilities to extensions.

To evaluate the two-dimensional intension of a statement S at a pair of worlds (V, W), one can ask: if V is actual, then if W were the case, would S be the case? This is reflected in some more natural English constructions: for example 'if water is XYZ, then water could not be H₂O'. To determine the truth-value of statements like these, one needs the full two-dimensional intension, as neither epistemic intension nor the subjunctive intension of the terms involved carries information about how to evaluate subjunctive statements under alternative epistemic possibilities. So for a fully general account of sentence's truth-values, one needs at least the full two-dimensional intension, although its full structure will be relevant only in rare cases.

Like an epistemic intension but unlike a subjunctive intension, a two-dimensional intension can be evaluated a priori. One needs no empirical information about the actual world, since all the relevant information is specified in the epistemic possibility. One might hold that the two-dimensional intension represents the true 'cognitive significance' of an expression, if one holds that the difference between the behavior of names and descriptions in counterfactual contexts represents a difference in cognitive significance; so one could hold that the two-dimensional intension is a Fregean 'sense' in some expanded understanding of the term. But we do not need to adjudicate that matter here.

An expression's actual subjunctive intension will correspond to the 'row' of the two-dimensional intension that results when the actual (centered) world is used as the epistemic possibility parameter; this cannot be evaluated a priori precisely because we do not know which epistemic possibility is actual. It is tempting to reconstruct an expression's epistemic intension as the 'diagonal' of the two-dimensional intension, which results when the same centered world is used as the epistemic possibility parameter and (in an uncentered version) as the metaphysical possibility parameter. This depends on just how the worlds are understood, however.

As before, there are two ways of understanding the worlds that function as epistemic possibilities. They can be understood as the same worlds that function as metaphysical possibilities, with the addition of a center. Or they can be understood as a separately defined class of worlds. The first option is more elegant, but requires a substantive philosophical theses about possibility. The second option is more complex, but it requires fewer philosophical commitments. A philosopher who holds that the existence of a god is metaphysically necessary but not a priori can embrace the second option but not the first, for example.

If one takes the second option: then one has two distinct classes of worlds, the epistemically possible worlds and the metaphysically possible worlds. These worlds may have certain relations to each other, but the two sets are nevertheless disjoint. One can evaluate a statement in an epistemically possible world, yielding the value of its epistemic intension there, and one can evaluate it in a metaphysically possible world, yielding the value of its subjunctive intension there. But because the spaces are distinct, there is no natural way to map epistemic possible worlds onto metaphysically possible worlds. As a result, one cannot reconstruct an epistemic intension as the 'diagonal' of the two-dimensional intension, and various other elegant properties are removed. This view also has the disadvantage of requiring a strong underlying modal dualism, with distinct modal primitives for each space of worlds.

If one takes the first option: then the worlds that function as epistemic possibilities are the same as the worlds that function as metaphysical possibilities, with the addition of a center. So epistemic intensions and metaphysical intensions are defined over (almost) the same space of worlds. To evaluate an expression's epistemic intension is evaluated at a world, we consider the world as an epistemic possibility: as a way our world might actually be. To evaluate an expression's subjunctive intension at a world, we consider the world as an explicitly counterfactual possibility: as a way our world might have been, but (probably) is not. One can say that in the first case, we consider the world as actual; in the second case, we consider the world as counterfactual.

The first option is compatible with all the familiar cases. Take the case of 'water', and consider a Twin Earth world where the watery stuff is XYZ (near the center, if required), and where H₂O is not watery. When we consider this world as actual, it is an instance of the epistemic possibility that water is XYZ: if W is actual, then water is XYZ. When we consider this world as counterfactual, it is an instance of the metaphysical possibility that water is H₂O: if W had been actual, then water would still have been H₂O, not XYZ. Something similar applies in all the familiar cases.

One might ask: is W *itself* a world where water is XYZ, or a world where water is H₂O? On a standard philosophical view, it is the latter, not the former. This is because the phrase 'a world in which S' is almost always read as invoking a world W in which the subjunctive intension of S is true, or for which S is true when W is considered as counterfactual. My own view is that this reading is arbitrary, and that the phrase 'a world in which S' is ambiguous between readings that invoke epistemic and subjunctive intensions. If so, there is no determinate answer to the question above. To remove the ambiguity one can define new locutions: for example, when the epistemic intension of S is true in W, then W is a world *at* which S; when the subjunctive intension of S is true in W, then W is a world *of* which S. In any case, this terminological issue does not matter too much for our purposes. As long as we are always clear about

how we are evaluating statements, and have clear conventions for understanding the relevant phrases, no confusion should result.

If the first option is accepted, various matters become more straightforward. Because there is only one space of worlds (apart from the difference involving centers), one can reconstruct an epistemic intension from the two-dimensional intension as a 'diagonal' intension, where the value of the epistemic intension at W is the value of the two-dimensional intension evaluate at W (considered as actual) and an uncentered version of W (considered as counterfactual). This view is also much simpler metaphysically, since it is compatible with a deep underlying modal monism, with just one space of worlds. Of course there is still a dualism of epistemic and metaphysical possibility in language, but this simply arises from the dual nature of semantic evaluation over a single space of worlds.

Given that any sentence S has an epistemic intension and a subjunctive intension, someone might ask: which of these is the *content* of S ? Which of these gives the *truth-conditions* of S ? What is the *propoposition* expressed by S ? My view is that we need not settle these questions one way or another. We can say that the epistemic intension is S 's *epistemic content* and that the subjunctive intension is S 's *subjunctive content*. As for the content of S (unqualified), this is a complex content that subsumes both of these and possibly more. Similarly, S has *epistemic truth-conditions* (showing how S 's truth depends on how the world turns out), and *subjunctive truth-conditions* (showing how S 's truth varies in counterfactual possibilities).

As for propositions, I have avoided this terminology, as it is multiply ambiguous. If one's conception of a proposition is a set of possible worlds (or something similar, such as a structure of intensions), then one could say that S expresses two propositions, an epistemic proposition and a subjunctive proposition. But if one's conception of a proposition is more generally of what remains semantically of S once the arbitrary clothing of a given language is stripped away, then one could say that S expresses a complex proposition with a two-dimensional structure. One should not run these two conceptions together: for example, the fact that an utterance of S expresses two propositions in the first sense in no way entails that the utterance is ambiguous, since ambiguity would involve expressing two propositions in the second sense. For my part, I prefer to use 'proposition' in the second, more general way; but I will largely avoid the expression here.

It is sometimes objected that the epistemic intension cannot be part of the content expressed by S at all, since it is really a matter of the content that S *would have* expressed had a token of S been present in a different context. This is a mistake. As we saw earlier, the epistemic intension of S can be evaluated in worlds that contain no token of S , and even if a world does contain such a token, its presence is usually irrelevant to evaluating an epistemic intension there. The epistemic intension of S is not defined in terms of counterfactual tokens of S at all. Rather, it is defined in terms of the first-order use of an *actual* expression S in evaluating various epistemic possibilities. This is precisely analogous to the way that the subjunctive intension of S is defined in terms of the first-order use of S in evaluating counterfactual possibilities. So epistemic intensions and subjunctive intensions are on a par here.

One could define the *contextual intension* of S as a function that maps centered worlds containing a token

of S at the center to the extension of that token in that world. This might resemble an epistemic intension in some ways, but it would not be an epistemic intension. First, an epistemic intension will be defined over many more worlds. Second, it is unclear just what it takes to be a token of S, and depending on what we require, the contextual intension may give very different results from an epistemic intension. If only orthographic properties are required, then there will be worlds where a token of 'water' refers to horses or to the number two. If the same extension is required, then a token of a name such as 'Hesperus' will pick out the same individual (Venus) in all worlds. Perhaps there is an intermediate requirement that gives roughly the same results as an epistemic intension, but this is not obvious. The obvious suggestion is to presuppose the notion of an epistemic intension, and require that a token of S have the same epistemic intension of S. However one does things, there is not much point in defining such a notion, since we have epistemic intensions to do the job already.

This matter bears on the two-dimensional semantic frameworks developed by Kaplan (1989) and Stalnaker (1974). The framework I have developed here resembles these in obvious respects, and it owes much to them. Epistemic intensions are analogous in certain respects to Stalnaker's "diagonal propositions" and to a version of Kaplan's "character", and subjunctive intensions are much the same as Stalnaker's "propositional content" and Kaplan's "content". But there are crucial differences. First: on Kaplan's and Stalnaker's frameworks, the analogs of epistemic intensions are defined in terms of the analogs of subjunctive intensions, whereas on the framework I have outlined, they are defined quite independently of subjunctive intensions. Second and most important: on Kaplan's and Stalnaker's frameworks, the worlds on the first dimension of evaluation are not considered as epistemic possibilities but as contexts of utterance, where evaluation requires the presence of a token of the expression within the context. As a result, this first dimension of evaluation yields a contextual intension, not an epistemic intension.

These features lead to strong limitations on using these frameworks for epistemic purposes, which both Kaplan and Stalnaker note. Kaplan's framework yields useful epistemic results only for indexicals and demonstratives, and not for names and natural kind terms. This is because the contextual intensions for indexicals and demonstratives behave much like their epistemic intensions, but the contextual intensions for names and natural kind terms behave very differently. (A name arguably picks out the same individual in every context, but it can apply to different individuals within epistemic possibilities.) Stalnaker (1999) argues that his "diagonal proposition" does not reflect matters of apriority directly, partly because of the problems involved in holding the meaning of an utterance constant across contexts. It seems that these frameworks are useful for epistemic purposes precisely to the degree that the notions involved resemble epistemic intensions.

A strong Fregean might criticize this framework from the other side, holding that only epistemic intensions, not subjunctive intensions, are part of the content of a sentence, and the subjunctive behavior can be accounted for by appealing to logical form or to a simpler semantic feature, as above. This matter is not cut and dried, but there are a number of advantages to including subjunctive intensions as an explicit semantic value.

First, subjunctive intensions allow a direct parallel between the treatment of apriority and of necessity,

and between epistemic and subjunctive evaluation. Second, subjunctive intensions give a more general account of subjunctive behavior than accounts that rely on actual extension, and consequently can more easily account for the distinctive subjunctive behavior of certain terms, such as 'cat'. Third, we have at least one content (the subjunctive intension) associated with an expression that is "objective" in the way that Frege required. Fourth, people have intuitions about "what is said" by an expression which at least sometimes better reflect subjunctive intensions than epistemic intensions (e.g. when I say 'I am Australian' and you say to me 'You are Australian', there is a sense in which we say the same thing, although our epistemic intensions differ). Fifth, the presence of subjunctive intensions allows for a degree of continuity with current philosophical frameworks in which a notion of "propositional content" closely tied to subjunctive intensions plays a central role, and allows a Fregean to retain some of the insights of this tradition rather than discarding them completely.

The addition of subjunctive intensions certainly goes beyond Frege's view. But it is a supplement to the view rather than a radical overhaul of the view. Senses, or epistemic intensions, are still present and playing the same role they always played. Sense and extension have merely been supplemented by a further semantic value in order to deal with various subjunctive and modal phenomena. Frege did not address these phenomena directly, so it is hard to know how he would have dealt with them. But I think this framework is compatible with the broad spirit of his views.

7 The epistemological argument

Kripke's second central argument against descriptive views of language is an epistemological argument. Kripke recognizes that a description theorist might accept that names are not modally equivalent to descriptions, but might nevertheless hold that names are epistemically equivalent to descriptions. So he raises some quite different considerations to argue that this is not the case. Again, while these arguments are most explicitly aimed at a description theorist, they are generally taken to have force against any broadly Fregean view.

The argument proceeds roughly as follows. The description theorist will hold that a name N (say, 'Gödel'), as used by a given speaker, is epistemically equivalent to some description or cluster of descriptions D (say, 'the person who proved the incompleteness of arithmetic'). On such a view, 'N (if it exists) is D' will be a priori for the speaker. But Kripke argues that no such sentence is a priori. In effect, he does this by arguing that for any description D, it is epistemically possible for the speaker that N is not D. If so, then N is not epistemically equivalent to D.

Take the name 'Gödel'. A speaker may associate a number of descriptions with her use of the name: 'the person who proved the incompleteness of arithmetic', and so on. But Kripke argues that there are (epistemically possible) scenarios in which the name 'Gödel' will turn out to refer to a person who does not satisfy any of the descriptions. So the speaker cannot know a priori that Gödel (if he exists) satisfies any of the descriptions.

One can put the argument strategy as follows. Take the name 'Gödel' and the description 'the man who

discovered the incompleteness of arithmetic'. Then there is an epistemically possible scenario in which the incompleteness of arithmetic was discovered by a man named 'Schmidt', and in which the proof was stolen and published by a man named 'Gödel', to whom the proof was thereafter attributed. If this scenario is actual, then the speaker's term 'Gödel' refers to the second man, not to the first. So it is epistemically possible for the speaker that Gödel did not prove the incompleteness of arithmetic, and there is no a priori equivalence between the name and the description.

Kripke argues that the same can be done for any description associated with the name (such as 'the man to whom the discovery of the incompleteness of arithmetic is commonly attributed'). There will always be epistemically possible scenarios such that if they are actual, the speaker's term 'Gödel' refers to someone who does not satisfy the description. (For example, a scenario in which the discovery is now commonly attributed to Hilbert, without the speaker realizing.) If this is right, then the name cannot be equivalent a priori to any description.

Kripke argues that there are actual cases in which most of the associated descriptions are false of the referent, as with the case of Peano, who did not discover the axioms associated with him, or with Jonah, who was probably not swallowed by a whale. So here it is not even true, let alone a priori, that the two are coextensive. And he also argues that there are some names such that the speaker has *no* associated description that could fix reference: a speaker can use the term 'Feynman' to refer to Feynman while knowing nothing more than that he is a famous physicist, for example, where that description is satisfied by many individuals. So again, it seems that reference is not fixed descriptively.

Does this argument against the description theory yield an argument against the intensional framework I have been outlining? It seems clear that it does not. This argument works with a conception of descriptions on which they correspond to linguistic expressions. When Kripke argues that the speaker the descriptions that the speaker "associates with" the name cannot fix reference, he always invokes linguistic descriptions that the speaker associates with the name, or at least explicit beliefs. But the intensional framework is not committed to the idea that descriptions always correspond to linguistic expressions; in fact, at least part of the motivation of the framework comes from an independent rejection of this idea. And the intensional framework is not even committed to the idea that the intensions associated with a name correspond to explicit beliefs of the speaker. So there is no clear argument against the intensional framework here.

In fact, Kripke's central method of argument seems to be obviously compatible with the intensional framework. A proponent of this framework could cast the argument strategy as follows. We want to show that for a given name *N* and description *D*, '*N* is *D*' is not a priori. To do this, we consider an epistemically possible world *W* with certain properties. We then reflect on a question such as the following: if *W* turns out to be actual, will it turn out that *N* is *D*? And we find that the answer is no. If so, the epistemic intension of '*N* is *D*' is false in *W*. So '*N* is *D*' is a priori.

On this interpretation, when we think about the Gödel/Schmidt case, for example, we are tacitly evaluating the epistemic intension of 'Gödel' at a world specified as in the example. When we consider that world as an epistemic possibility, it reveals itself as an instance of the epistemic possibility that

Gödel did not discover incompleteness. That is, we find that the epistemic intension of 'Gödel' does not pick out the prover in this world; it picks out the publisher. If so, the epistemic intensions of 'Gödel' and of 'the man who discovered the incompleteness of arithmetic' are distinct.

In a way, what is going on here is analogous to what goes on in the analysis of a term such as 'knowledge', as discussed above. Someone might hold that 'knowledge' is equivalent a priori to some description, such as 'justified true belief'. But then we come up with a scenario such that in that scenario, something falls within the extension of 'justified true belief' but not of 'knowledge'. So we conclude that the two are not equivalent a priori. The process repeats itself for other descriptions, suggesting that 'knowledge' is not a priori equivalent to any such description. This suggests that the intension of 'knowledge' cannot be precisely captured in a linguistic description.

In a similar way, Kripke's arguments suggest that the epistemic intension of a name such as 'Gödel' cannot be precisely captured in a linguistic description. But they do nothing to suggest that the epistemic intension does not exist. And the epistemic intension still mirrors the cognitive significance of the name. The identity 'N is D' is a posteriori precisely because the two expressions have different epistemic intensions; that is, precisely because there is an epistemic possibility where they come apart. And the subject has the ability to evaluate the name's referent within this epistemic possibility, just as can be done with expressions in general.

The intension of an expression such as 'knowledge' can at least be approximated by certain linguistic descriptions, such as 'justified true belief', and by longer and longer versions that come gradually closer to the true intension. One might wonder whether something similar can be done with a name such as 'Gödel'. Can the epistemic intension of the name in the case above at least be approximated by a linguistic description? This is not compulsory for the intentionals framework, but it can at least be enlightening to look. A side benefit is that it provides some sort of at least approximate account of the features of the world in virtue of which the epistemic intension applies.

To answer this question, one needs to consider: when speakers use a name such as 'Gödel' or 'Feynman' in cases such as those above, how do they determine the referent of the name, given sufficient information about the world? For example, if someone knows only that Feynman is a famous physicist and that Gell-Mann is a famous physicist, how will external information allow her to identify the distinct referents of 'Feynman' and 'Gell-Mann'? The answer seems clear: she will look to *others'* use of the name. Further information will allow her to determine that members of their community use 'Feynman' to refer to a certain individual, and that they use 'Gell-Mann' to refer to a different individual. Once she has this information, she will have no problem determining that her own use of 'Feynman' refers to the first, and that her own use of 'Gell-Mann' refers to the second.

This suggests if we want to approximate the epistemic intension of the speaker's use of 'Feynman' in a description, one might start with something like 'the person called 'Feynman' by those from whom I acquired the name'.[*] It certainly seems that if relevant information about others' uses is specified in an epistemic possibility, then this sort of description will usually give the right results. The same goes for the 'Gödel' epistemic possibility. In all these cases, it seems that a name is being used *deferentially*: in using

a name, the speaker defers to others who use the name. So maybe the description above is at least a good first approximation.[*]

*[[Kripke considers some potential descriptions in this vicinity. He discusses 'the man Jones calls 'Gödel' (1980, p. 92), where the speaker believes he acquired the name from Jones, and dismisses it on the ground that the belief may be false. And he discusses 'the man called 'Glumph' from the people from whom I got it (whoever they are), provided that my present determination of the reference satisfies the conditions sketched in *Naming and Necessity* and whatever other conditions need be satisfied' (1980, p. 162), and dismisses it on grounds connected to the second part of the description. Strangely, he never considers the obvious intermediate description that avoids both of these problems.]]

*[[Replies to Kripke's epistemic arguments that appeal to metalinguistic descriptions of this general sort are given by Fumerton 1989, Jackson 1998, and Lewis 1986, among others.]]

There are two sorts of objections that might be made to this sort of description. First, it might be held that it does not always give the right results. For example, it may be epistemically possible that the speaker misheard or misremembered the name, and that others were really using the name 'Fireman'. In such an epistemic possibility, the description above will give the wrong results. But this is just the sort of thing that we should expect, given the imperfection of descriptions. As with 'knowledge', we could try to move to closer approximations. Perhaps 'The referent of the relevant name used by the person from whom I acquired the antecedent of my current term 'Gödel' would do a better job. But no doubt there would be further counterexamples, just as with 'knowledge'. But as in all these cases, the most this shows is that any such approximation is imperfect. One refutes these approximations by evaluating the epistemic intension in certain epistemic possibilities and showing that the approximation gives the wrong results; so this sort of argument does nothing to show that the epistemic intension does not exist.

Second, it might be held that this sort of description is "circular". perhaps because it appeals to the notion of reference. The circularity is not obvious, however. If one appealed to the notion of reference in a definition of 'reference', there would be a danger of circularity. But a definition of 'reference' is not being offered here. If the descriptivist were to offer a definition of 'reference' it might be something like 'whatever satisfies a canonical description D associated with N'. This is a general, uncircular definition with no appeal to reference (except in the notion of satisfying a description, which presumably is to be accounted for separately). At worst there is a danger that if a description D involves 'reference', it will not yield a determinate result, since evaluating the description will require evaluating reference, which will require evaluating another description, which will require evaluating reference, and so on. But a descriptivist can reply that this is merely a recursive situation, not a circular one, and that the process will always eventually be grounded in a use of the name whose associated description does not involve 'reference' at all, such as an initial baptism. The descriptivist might even hold that this is a natural way to capture the insights of the causal theory of reference in a descriptivist framework: reference proceeds through recursive deference to others, ultimately grounded in an initial baptism.

In any case, this worry does not arise on the intensional framework. The epistemic intension of a name is simply a function from worlds to individuals that reflects a rational ability to determine a specific individual in a given epistemic possibility. It is not a description, and so makes no use of 'reference'.

Perhaps one might worry that if something like the picture above is right, then *evaluating* the epistemic intension of a term like 'Gödel' at a world will require having explicit information about others' reference within that world, and that this would be circular. It is not clear why this would be circular, but in any case, explicit information about reference would not be required. Information about the epistemic intensions that others associate with various names would suffice (along with other information about causal relations, the properties of various objects, and so on). This information might itself be derivable from information about other mental and/or physical states, or it might be some sort of mental primitive; that does not matter. With information about others' epistemic intensions, reference will be determined. And if there is a worry about *evaluating* others' epistemic intensions given that they are deferential also, then knowledge of epistemic intensions across a whole community (including its history) will suffice, since the deference will ultimately be grounded in a nondeferential use.

So given relevant information about the physical and mental states of individuals in the community, there will be no problem evaluating the epistemic intension. And the use of this information is in no way ad hoc; it corresponds to the information that we use in evaluating reference across various epistemic possibilities. In particular, when a name is used deferentially, information about the linguistic and cognitive practices of others will always be relevant.

Of course not every use of a name is a deferential use, so not every epistemic intension will function in this way. When a name is introduced in an initial baptism, there will be no deferential element involved; to evaluate reference in the actual world (and across epistemic possibilities), the speaker will not usually need information about the cognitive states of others. The same goes for some names used for very familiar referents. Say that a wife uses the name 'Fred' for her husband and has done so for years. In such a case, even were the speaker to discover (to her surprise) that no-one else in the community used that name for her husband, she would still reasonably hold that the name she uses refers to the spouse. If such a situation turned out to be actual, her utterance 'Fred is my husband' would plausibly be true, not false. This suggests that the epistemic intension of her use of the name has no deferential element.

There are also many intermediate cases, where a name is used with some mixture of deferential and nondeferential elements, so that for a speaker to determine reference of the name, relevant information will include both information about others' usage and independent information about properties of the referent (perhaps corresponding to some of the speakers' beliefs involving the name). This sort of intermediate case will be necessary to account for cases such as 'Madagascar' (Evans 1977), where the referent of our use of the name (an island) differs from the referent of the usual use (part of mainland Africa). If every use since the initial baptism was entirely deferential, this could not happen. So some uses in the causal chain must have been not entirely deferential, with epistemic intensions that were partly influenced by a speaker's beliefs. It is easy to imagine that even if the beliefs have only a small influence on a given speaker's epistemic intension, the effect of this influence would amplify as a causal chain proceeds. If every speaker in the chain has a small component of influence from the belief, then deference to a speaker whose epistemic intension is also influenced by the belief will increase the effect. In the case of 'Madagascar', the result might be that uses of the name would initially refer to the mainland location, would proceed through a period of divided reference, and would eventually emerge as referring to the island.

(One other subtlety: to evaluate a deferential use of the name in an epistemic possibility, the speaker may need the name itself (as used by her) to be present in an epistemic possibility, so that she can determine where that name was acquired. This suggests that deferential uses are exceptions to the principle that tokens of an expression need not be present within an epistemic possibility. I think that this sort of case is best handled by having one or more optional marked thoughts ('this thought') present at the center of a world, which be used to trace deferential reference in an epistemic possibility.)

Kripke also gives an epistemological argument against descriptive views of natural-kind terms (1980, pp. 116-23). He argues that it is not a priori that gold is yellow, or that tigers are striped, and so on, in effect because there are epistemic possibilities in which these statements could turn out to be false (if we were suffering from various illusions, for example). In a similar way, Putnam (1975) argues that it is not a priori that cats are animals, since there are epistemic possibilities in which it would turn out that they are robots from Mars. As before, these arguments do not apply to the intensional account: they proceed by evaluating an epistemic intension at various epistemic possibilities, and they show at most that the epistemic intensions is not equivalent to the relevant descriptions. They also suggest that a causal link between the term and the referent plays an important role; so any better descriptive approximation of the intension should give a significant role to this link.

I conclude that the epistemological arguments have no force against an intensional account of Fregean sense.

8 The argument from variability

Another argument against Fregean view is rarely articulated explicitly,[*] but it may lie in the background of some opposition to the Fregean view. This argument notes that different speakers may associate *different* cognitive significance with the same name. For example, an identity such as "Bill Smith is William Smith" may be cognitively significant for one speaker (e.g. one who has heard of the same person under two names in two different contexts), and cognitively insignificant for another (e.g. the person's partner, who uses both names indiscriminately). So if Fregean sense is to reflect cognitive significance, then the sense of a name must vary between speakers. But if the sense of a name can vary between speakers, it is not part of the *meaning* of the name at all. The same may apply to other expressions, such as natural kind terms: the cognitive significance of such a term is variable, so its sense must be variable, so its sense cannot be part of its meaning.

*[[Frege himself (19xx) addresses objections based on variability, in the context of indexicals rather than names. Burge (1979) stresses that Frege's conception of sense differs from contemporary conceptions of linguistic meaning, in part because of variability.]]

We need not spend much time on the first part of this argument. On the account I have given, it is clear that the epistemic intension of a name can vary between speakers. We saw this in the discussion just concluded: the first user of a name may use the name with one epistemic intension, and later users may use it with a quite different epistemic intension. When Leverrier introduced the term 'Neptune' as a term

for whatever planet was perturbing the orbit of Uranus, then the epistemic intension of his use of the term functioned roughly as described. But the next speaker - perhaps his wife, who knew only that Neptune was an astronomical object for which her husband was searching - might have used it with a different epistemic intension. And later users might well use it in a deferential way (as with 'Feynman' above), with an epistemic intension that reflects this. So different speakers can clearly have different epistemic intensions for the same name.

The same might apply to natural kind terms, though this is not quite so clear. For example, two speakers might have been exposed to different forms of water: one has only been exposed to water in liquid form (knowing nothing of a solid form), and the other has been exposed only to water in solid form (knowing nothing of a liquid form). It might be that for the first speaker, the epistemic intension of 'water' might function to pick out (roughly) a substance that takes on a certain liquid form, and that for the second speaker, the epistemic intension of 'water' might function to pick out a substance that takes on a certain solid form. In the actual world, these epistemic intensions both pick out the same substance, but in other epistemic possibilities, their extensions will differ. But arguably both are using the same word 'water'. If so, there is no epistemic intension for 'water' that is common across all users of the term.

The same applies in an even stronger way to demonstratives such as 'that'. Here different uses of the same term by the same speaker can have different epistemic intensions. It is plausible that the epistemic intension of 'that' depends at least in part on a speaker's intentions, which may differ between uses of the term: on one occasion, the speaker may intend to refer to an object on her left, and on another occasion, to an object on her right. If so, the epistemic intensions on these different uses may differ. So there is no epistemic intension for 'that' that is common across all uses of the term, even for a single speaker.

This suggests that for a general account, epistemic intensions cannot be assigned to expression *types* but rather must be assigned to expression *tokens*. There will be some terms for which epistemic intensions are constant across all tokens of an expression type - some descriptive terms and indexicals, for example - in which case an epistemic intension can also be assigned to the type. And for names and natural kind terms, epistemic intensions might be constant at least across an individual speaker's use of an expression (at least within a limited time frame), so there could be an assignment to more limited types. But a fully general account requires that epistemic intensions are assigned to tokens.[*]

*[[It should be noted that this difference in epistemic intensions between speakers goes along with differences in which sentences are a priori for a given speaker. For example, 'Neptune (if it exists) affects the orbit of Uranus' may be a priori for Leverrier, but not for his wife. Leverrier could use the sentence to express a priori knowledge, but his wife could not. So insofar as the apriority of various sentences is invoked in the analysis of epistemic intensions, it will generally be a speaker-relative notion that is relevant. (This is reflected in Kripke's own discussion (e.g. 1980, p. 73) which usually talks of a sentence being "a priori for a speaker".) Of course the notion is still distinct from the notion of what the speaker knows a priori; it is a notion of what is knowable a priori for a speaker, given ideal rational reflection.]]

What of the second part of the argument: that if epistemic intensions can vary between tokens of a type, then they are not meanings? This issue is largely terminological. If it is stipulated that meanings are constant across all tokens of a type, then epistemic intensions are not meanings. If this is not stipulated,

then epistemic intensions might be meanings. We could distinguish 'type meanings' and 'token meanings', and allow that epistemic intensions are not (in the general case) type meanings, but they are token meanings. Or we can use a different term, such as 'content', for the sort of meanings that can vary between tokens of an expression type. It is not clear that a substantive issue remains once the terminological issue is cleared up.

It might be insisted that if epistemic intensions can vary between tokens of a type, then they are not part of *language*, they are not an aspect of *linguistic* content, and that perhaps they do not fall within the domain of the philosophy of language at all. Again, this is a terminological issue. One might say that the epistemic intensions of names are not part of "a language" such as English, where this is considered as what is common between all English speakers. But this is no reason to deny that they are part of language in a broader sense, and they are in the domain of the philosophy of language. Of course these terminological issues are largely sociological in origin; so if someone resists on all these issues, it may help to point to some historical examples.

First, Frege. As we have seen, Frege himself held that the sense of an expression could vary between tokens of a type, in the case of indexicals and names. Frege's theory of sense is generally taken to be one of the most important theories in the philosophy of language, and Frege himself is perhaps the most important figure in the field. It would be odd to hold that the theory of sense (as sense was characterized on Frege's account) has nothing to do with language, and that Frege's theory is not part of the philosophy of language.

Second, Kripke. It is striking that in all his arguments against the Fregean view, Kripke never mentions an argument from variability. He appears to take it as obvious that on a Fregean theory, the descriptive content of a name will vary between speakers (his relevant formulations are all speaker-relative), and does not mention this as an objection. Given the number of other objections that are developed against a Fregean account, this suggests that Kripke took variability to be no objection to sense as an aspect of language. And it suggests that if the worst problem for a Fregean is variability, then a Fregean view of language would be broadly correct.

Third, Kaplan. Kaplan's theory of character and content (for indexicals and demonstratives) is generally taken to be one of the most important theories in the philosophy of language in recent years. On Kaplan's account, it is clear that both the character and the content of a demonstrative such as 'that' will vary between uses of the term.[*] So if epistemic intensions are not an aspect of language, then neither are character or content.

*[[Kaplan's formal language for demonstratives deals with this issue by holding that tokens of demonstratives that are associated with different intentions or demonstrations are in fact tokens of a different word. But it is extremely plausible that in a natural language such as English, different tokens of a demonstrative such as 'that' are tokens of the same word; so this formal stipulation still leaves character distinct from linguistic meaning in natural language. See Braun 19xx for critical discussions, and for proposals regarding the linguistic meaning of demonstratives.]]

Someone might object that the epistemic intension of an utterance is not always part of what that

utterance *communicates*, since a speaker and a hearer may associate different epistemic intensions with the same expression, and the hearer may not know what the speaker's epistemic intension is. It is true for that this reason, epistemic intensions are not always communicated: a speaker's utterance with a given epistemic intension may cause the hearer to acquire a belief with a different epistemic intension. But it is not clear why this sort of communicative property is a *sine qua non* for any aspect of the content of language. Certainly it is not satisfied by Frege's sense or by Kaplan's character or content, for example. At worst, epistemic intensions are in the same boat as these paradigmatic aspects of linguistic content.

It might also be objected that epistemic intensions are an aspect of "speaker meaning" rather than "semantic meaning". This objection exploits a distinction Kripke (19xx) exploited to deal with Donnellan's (19xx) distinction between the "referential" and "attributive" use of expressions such as 'the man in the corner drinking champagne', used by a speaker intended to refer to a man he is looking at in the corner. If the man in the corner is actually drinking a martini, there is a sense in which the expression refers to him (the referential sense), and a sense in which it does not (the attributive sense). Kripke argues that the attributive reading reflects the "semantic reference" of the expression, while the referential reading is merely an aspect of "speaker's reference". Perhaps something similar applies to epistemic intensions?

I think Kripke's analysis of Donnellan's cases is plausible, but it does not generalize to the case of epistemic intensions. As Kripke says:

The semantic referent of a designator is given by a *general* intention of the speaker to refer to a certain object whenever the designator is used. The speaker's referent is given by a *specific* intention, on a given occasion, to refer to a certain object.

It is clear that on this definition, the epistemic intension of a name or natural kind term is more akin to semantic reference than to speaker's reference, since it reflects a general intention on the speaker's part, not a specific intention. Indeed, the epistemic intension of 'the man in the corner drinking champagne' in the case above plausibly picks out the semantic referent (no-one), not the speaker's referent (the man in the corner). One might, if one wished, introduce a corresponding difference between "semantic epistemic intension" and "speaker's epistemic intension", where the latter picks out the man who is ostended irrespective of his other properties. But even then, epistemic intensions would be in the same boat as reference, and there would be a clear notion that falls on Kripke's "semantic" side of things.

Perhaps there is some *other* distinction that might be drawn between something one might call "semantic meaning" and something one might call "speaker meaning": for example, if one stipulates that semantic meaning must be constant across all tokens of a type, while speaker meaning can vary between speakers. But that would be a very different distinction from Kripke's, and it would do nothing to suggest that epistemic intensions are in the same boat as Donnellan's 'referential' uses of descriptions.

A final objection might be that epistemic intensions are not part of language, since they derive entirely from the contents of thought. On this view, it is the speaker's *concept* of water that has an epistemic intension, and it is the conceptual content that varies between speakers, not any sort of linguistic content.

In response: one can accept that epistemic intensions are associated with concepts and thoughts,[*] and even that the epistemic intensions of linguistic expressions are derivative in some way on the epistemic intensions of thoughts, but this not entail that epistemic intensions are not also part of the content of language. On my view, all linguistic content is derivative in some way on mental content; this applies just as much to expressions whose content is constant between speakers as to expressions whose content varies. But there is no reason to hold that by definition, linguistic content cannot derive from mental content. If this were so, then my view would entail that there is no linguistic content. Perhaps my view is wrong, but if so, this is a substantive point, not a terminological one. It is also worth noting that Kripke's own discussion above suggests that linguistic content can derive from mental content (a speaker's intentions). So I think that there is no objection to epistemic intension as a sort of linguistic content here.

*[[See "The Components of Content", which uses epistemic intensions to give an account of the "narrow content" of thought.]]

Ultimately, the best way to deal with any terminological issue is to reflect on the use to which a term is being put, and to determine which sense is most relevant to a given purpose. It may be that there are some purposes, the most relevant notion of the "meaning" or "content" of linguistic expressions is one on which meaning and content are required to be constant across all tokens of a type. Such purposes might include those of determining what is built into the semantic structure of a language such as English, giving an account of what is required to competently use an expression of a given type, and perhaps addressing certain questions about what an expression will communicate between arbitrary speakers of a language. For these purposes, one can invoke a notion of meaning on which universality is required. On this conception of meaning, epistemic intensions may be part of the meanings of indexicals and some descriptive terms, but they will not be part of the meanings of names and natural kind terms.

For many other purposes, we do not need such a narrow notion of meaning or content, and we often need a broader notion. This includes most of the uses to which the philosophy of language and notions of meaning are put in other areas of philosophy, as well as questions within the philosophy of language itself.

One example: notions of meaning and content are often taken to be central in analyzing questions about necessity and possibility, which in turn play a crucial role in analyzing many metaphysical issues. But for this role, the question of whether meaning or content is constant across speakers is almost entirely irrelevant. It would make very little difference to the deepest issues if there were just one speaker of a language, or if different speakers used terms with different meanings or content. If I use an epistemic intension to reach a metaphysical conclusion about water, the worst possible consequence of variability will be that someone else will not be able to reach a similar conclusion using their term "water". This seems unlikely to happen, due to the common referent. But even if it did, it would do nothing to invalidate my conclusion; it would just mean that someone else would have to express the conclusion differently. Once terminological issues are cleared up (as they often need to be), the substantive points will be as before. So if epistemic intensions are otherwise relevant to answering these questions, as I think they are, it is no objection that they can vary between speakers.

The same goes for many or most applications of the philosophy of language in metaphysics, epistemology, the philosophy of mind, the philosophy of science, and other areas. Some uses to which a notion of meaning or content is naturally put include: questions concerning the rational role of language; questions concerning the connection between language and thought; questions concerning the evolution of meaning as scientific theories develop; questions about scientific explanation; questions concerning the semantics of belief ascription and indicative conditionals; questions about how reference is fixed; questions about the a priori; and so on. For most of these purposes, a requirement that meaning or content is constant over all tokens of an expression type is at best irrelevant, and is at worst harmful. So objections based on variability between speakers will be largely irrelevant to the use of notions of meaning and content in these areas.

9 Conclusion

I have argued that a broadly Fregean account of meaning is tenable. On this account, the notion of an epistemic intension plays the role of a Fregean notion of sense. Epistemic intensions are not the same as Fregean sense in all respects, but they are similar in many respects, and they allow versions of the core Fregean requirements on sense to be satisfied. It may be useful to summarize where the core Fregean theses stand in light of the preceding discussion.

Thesis (1), that every expression has a sense, has been preserved. It has also been augmented by a thesis holding that every expression has a subjunctive intension in addition to its sense (its epistemic intension).

Thesis (2), that sense reflects cognitive significance has been preserved in a slightly modified form. In the modified form, cognitive significance is understood as non-apriority. So when an identity is not a priori, the expressions involved have different senses; when two sentences are not equivalent a priori, they have different senses; and so on.

Thesis (3), that the sense of a complex expression depends on the sense of its parts, has been preserved in a modified form. The thesis as it stands holds for all expressions except those involving modal and subjunctive contexts. To handle such expressions, one needs a slightly modified thesis: that the sense of an expression depends on the semantic values of its parts, where the semantic value may include elements (such as subjunctive intension) that go beyond sense. It is also the case that all aspects of semantic value (extension, epistemic intension, subjunctive intension) of a complex expression depends on the semantic values of the parts.

Thesis (4), that sense determines extension, has been preserved in the version that holds that sense determines extension in combination with the world.

Thesis (5), that the sense of a sentence has an absolute truth-value, has been discarded. The sense of a sentence has a truth-value only relative to a subject and a time. Thesis (6), that expressions refer to their senses in indirect contexts, has also been discarded. A more complex account of these contexts is still required. Thesis (7), that the sense of an expression can vary between speakers and between occasions of

use has been preserved.

I have also argued that the most common objections to Fregean theories can be handled by such an account. At most, these objections show that (i) senses are indexical, (ii) senses should be supplemented by a further semantic value, a subjunctive intension, (iii) senses should be understood as intensions, not descriptions, and (iv) the sense of an expression can vary between speakers and between occasions of use.

I have argued that extension, epistemic intension, and subjunctive intension are all part of the meaning and content of an expression, but I have not argued that these exhaust the meaning or content of an expression. In fact I think that they do not. First, there are plausibly aspects of meaning that have nothing to do with the determination of truth: the difference between 'and' and 'but' is an example. Second, there may be expressions that are a priori equivalent to each other, but that nevertheless have different meaning due to some more fine-grained cognitive difference: the difference between 'equilateral triangle' and 'equiangular triangle' is an example. To handle this last sort of difference, I think that one may need senses of a variety that are more fine-grained than epistemic intensions. The notion of an epistemic intension might be extended to do this (by moving to a more fine-grained space of epistemic possibilities), but this is a separate story. It may also be that for some purposes, the meanings or contents of complex expressions may need to be taken as structured complexes of extensions and/or intensions; this is entirely compatible with the framework I have outlined.

I have also not given a conclusive demonstration that epistemic intensions of expressions exist and have all the properties I have attributed to them. A conclusive demonstration would require some more precise definitions, and a rebuttal of all counterarguments. I have argued that there is a strong prima facie case that epistemic intensions exist and have the properties I have attributed to them, however, and I have argued that the most obvious counterarguments can be rebutted. It may be that there are other arguments against the view; if so, I would be very interested to hear them. In the meantime, I think that a broadly Fregean approach to meaning holds considerable promise.

The Nature of Epistemic Space

David J. Chalmers

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

[N.B. A very rough draft, with no notes or citations yet, and almost certainly lots of mistakes.]

1 Ways the world might be

There are many ways the world might be, for all I know. For all I know, it might be that there is life on Jupiter, and it might be that there is not. It might be that Australia will win the next Ashes series, and it might be that they will not. It might be that my great-grandfather was my great-grandmother's second cousin, and it might be that he was not. It might be that copper is a compound, and it might be that it is not.

There are even more ways the world might be, for all I know with certainty. It might be that there are three chairs in this room, and it might be that there are not. It might be that water is H₂O, and it might be that it is not. It might be that my father was born in Egypt, and it might be that he was not. It might be that I have a body, and it might be that I do not.

We can say that it is *epistemically possible* for a subject that P, when it might be that P for all the subject knows. So it is epistemically possible for me that there is life on Jupiter, or that copper is a compound. One can define various different standards of epistemic possibility, corresponding to various different standards for knowledge. For example, one might say that it is *epistemically possible in the Cartesian sense* (for a subject) that P when it might be that P, for all a subject knows with certainty. So in the Cartesian sense, it is epistemically possible for me that water is not H₂O, and it is epistemically possible for me that I do not have a body.

A natural way to think about epistemic possibility is as follows. When it is epistemically possible (for a subject) that P, there is an epistemically possible *scenario* (for that subject) in which P. We can think of a scenario as a maximally specific way things might be: a sort of epistemically possible world, in a loose and intuitive sense. On this picture, corresponding to the epistemic possibility that Australia will win the next Ashes series are various epistemically possible scenarios in which they win in all sorts of different ways. And corresponding to the Cartesian epistemic possibility that I have no body are various scenarios in which I am disembodied, each epistemically possible by the Cartesian standard: e.g. scenarios in which I am a brain in a vat, or in which I am a disembodied Cartesian mind.

To fill out this picture a bit further, we might imagine that there is an overarching space of scenarios. We can think of these scenarios as constituting *epistemic space*. If a subject did not know anything, all scenarios would be epistemically possible for the subject. When a subject knows something, some scenarios are excluded. Every piece of substantive knowledge corresponds to a division in epistemic space: some scenarios are excluded out as epistemically impossible for the subject, while others are left open. More specifically, it is natural to hold that for a given P , there may be scenarios in which P is the case, and scenarios in which P is not the case. Then when a subject knows that P , scenarios in which P is not the case are excluded, while others are left open. The scenarios that are epistemically possible for a subject are those that are not excluded by any knowledge of the subject.

One can naturally suppose that the space of scenarios is equally divided by *belief*, and perhaps that the division by belief underlies the division by knowledge. Every substantive belief, whether or not it qualifies as knowledge, corresponds to a division in the space of scenarios. When a subject believes that P , we might say that some scenarios (in particular, scenarios in which not- P) are ruled out as *doxastically impossible*, while others are left open. A scenario is doxastically possible for a subject if and only if it is not doxastically ruled out by any of the subject's beliefs. When a belief qualifies as knowledge, the scenarios ruled out as doxastically impossible are also ruled out as epistemically impossible.

A picture of this general sort is often present in philosophical discussions of knowledge and belief. Within epistemology, it is common to think of knowledge in terms of the "elimination of possibilities", with some sort of underlying space of possibilities presumed. In discussions of skepticism, for example, the fact that certain skeptical scenarios are not eliminated is used as evidence that certain knowledge claims are not true. And in epistemic logic and the theory of belief revision, it is common to model epistemic possibility using epistemic relations to an underlying space of possible worlds.

It is surprisingly difficult, however, to make the intuitive picture precise. What sort of possibilities we are dealing with here? In particular, what is a scenario? And what is the relationship between scenarios and items of knowledge and belief?

It is natural to think of scenarios as possible worlds, and to think of a scenario in which P as a world in which P . But it is immediately clear that this will not work, at least on the most common contemporary understanding of possible worlds. There are subjects for whom it is epistemically possible that Hesperus is not Phosphorus; but on the usual understanding, there is no possible world in which Hesperus is not Phosphorus. It is epistemically possible for me that my great-grandparents were cousins and it is epistemically possible that they were not; but on the usual understanding, my great-grandparents are cousins either in all worlds in which they exist or in none. In the Cartesian sense, it is epistemically possible for me that water is not H_2O , but on the usual understanding (assuming that water really is H_2O), there are no possible worlds in which water is not H_2O . So if we are to maintain that it is epistemically possible that P iff there is an epistemically possible scenario in which P , we cannot identify a scenario in which P with a possible world in which P , at least on the usual understanding.

Some might react to this by denying the intuitions about what is epistemically possible (e.g. holding that it is never epistemically possible that Hesperus is not Phosphorus), and some might react by denying the coherence of the picture connecting epistemic possibility to epistemically possible scenarios. I think that both reactions would be premature: the first loses touch with the phenomenon we are trying to analyze, and the second assumes that possible worlds as currently understood are the only available tool.

Instead, we should try to understand epistemic possibility on its own terms. We are not dealing here with counterfactual space: the space of ways the world might have been. Here, we are dealing with epistemic space: the space of ways the world might be. This epistemic space calls for its own epistemic tools of analysis. Where the analysis of counterfactual space invokes possible worlds as maximally specific ways the world might have been, the analysis of epistemic space should invoke scenarios as maximally specific ways the world might be. The two notions are quite distinct, although they have a deep underlying relationship.

In this paper, I will explore some ways of understanding epistemic space. I hope to make at least a *prima facie* case that a version of the notion is coherent, and that there are no decisive problems with it. And I will argue that the notion, if coherent, has a large philosophical payoff in the analysis of the content of thought and the meaning of language. It promises an analysis of the dimensions of meaning and content that are tied constitutively to the epistemic dimension. In particular, it promises an account of a sort of Fregean sense for linguistic expressions, and it promises an account of the narrow content of thoughts.

2 Preliminaries

On the picture suggested above, we might say that the notion of *strict epistemic possibility* - ways the world might be, for all we know - is undergirded by a notion of *deep epistemic possibility* - ways the world might be, prior to what anyone knows. Unlike strict epistemic possibility, deep epistemic possibility does not depend on a particular state of knowledge, and is not relative to a subject. Whereas it is strictly epistemically possible (for a subject) that P when there is some epistemically possible scenario (for that subject) in which P, it is deeply epistemically possible that P when there is some deeply epistemically possible scenario in which P. Since all scenarios are deeply epistemically possible on this picture, we can put this more simply: it is deeply epistemically possible that P when there is some scenario in which P.

How can we understand deep epistemic possibility more precisely? A problem immediately presents itself. It might be held that for *any* proposition P, there is some subject for whom P is strictly epistemically possible. Perhaps there are subjects who are so limited or so confused that they know nothing at all: if so, then any P is epistemically possible for them. Or for any given P, there might be a subject who is so limited or so confused with respect to the domain of P that they know nothing in this domain: if so, it is plausible that P will be epistemically possible for the subject. This may apply even to elementary logical contradictions, and to the negation of trivialities. If so, then if strict epistemic possibility entails deep epistemic possibility, then for all P, it will be deeply epistemically possible that P.

This notion of deep epistemic possibility might be useful for some purposes, but it is not very useful for our purposes. If it is deeply epistemically possible that P for all P , so that there is a scenario in which P for all P , then the structure of epistemic space will be near-trivial. Under certain natural assumptions (of a kind outlined below), for any propositions $P, Q, R, \dots$, there will be distinct scenarios in which $P \& Q \& R \& \dots$, $P \& Q \& \sim R \& \dots$, $P \& \sim Q \& R \& \dots$, and so on (including all sorts of scenarios in which $P \& \sim P$, and so on). So the space of scenarios will be as fine-grained as the power set of the space of propositions. This would rob the space of scenarios of interesting structure, and it would also trivialize the relationship between scenarios and belief. Beliefs and knowledge would divide the space of scenarios only in a trivial way, and there could be no nontrivial inferences from uneliminated scenarios to contents of belief or knowledge, or vice versa.

For a more useful notion, we need a more constrained notion of epistemic possibility, one that builds in some degree of rationality, and precludes incoherent possibilities such as those in which $P \& \sim P$. There are various ways that one might do this, but the most obvious way is to exploit the idea that these incoherent possibilities can be ruled out a priori. We can say that P is deeply epistemically possible when it is not a priori that $\sim P$. On this understanding, if P is a priori false, there will be no scenario in which P . Scenarios will now correspond to maximally specific *coherent* ways the world might be. We might think of the resulting notion of epistemic space as *ideal epistemic space*, or *rational space*.

It is clear that the idealized notion diverges from the ordinary understanding of epistemic possibility, at least within the a priori domain. For example, on the ordinary understanding, negations of complex mathematical and logical theorems can be epistemically possible (I do not know enough to rule them out), but on the idealized understanding above, these negations will be deeply epistemically impossible. Ideal epistemic space is well-suited to analyzing much empirical knowledge and belief, however, and it also yields an elegant system with many useful properties. So I will focus on the idealized notion of deep epistemic possibility and of epistemic space, at least for now. Later, I will discuss ways in which the idealization might be relaxed.

Another question arises: what are the objects of deep epistemic possibility, and what sort of objects are true or false at scenarios? So far I have often talked as if these are propositions. I think that this view is ultimately correct, but the contested nature of propositions raises difficulties in developing the approach. For example, on some theoretical views of propositions (e.g. views involving singular propositions, and some views involving sets of possible worlds), "Hesperus is Hesperus" and "Hesperus is Phosphorus" express the same proposition. But on the face of it, it is deeply epistemically possible that Hesperus is not Phosphorus (we cannot know a priori that Hesperus is Phosphorus), but it is not deeply epistemically possible that Hesperus is not Hesperus (we can know a priori that Hesperus is Hesperus). So it is not easy for these views of propositions to capture this difference: they must either deny the epistemic possibility claim, or accept that the epistemic possibility that P is not to be analyzed as the epistemic possibility of the proposition that P . I think that there is an alternative understanding of propositions that avoids these problems, and that the approach developed in this paper can itself help support such an understanding, but it would be problematic to assume such an understanding at the outset. To avoid this sort of theory-dependent confusion, it is best to start without talking about propositions at all.

Instead, we can associate epistemic possibility, and truth or falsity at scenarios, with something in the vicinity of *beliefs* and *sentences*. I will take beliefs to be token mental states: occurrent beliefs will be the paradigm example. The sentences I will be concerned with are sentence tokens that are used to make assertions: call such a token a *statement*. I will take it that there is a relation of expression between statements and beliefs, and that when a statement expresses a belief, the statement and the belief have the same truth-value. Statements usually express a belief of the subject making the statement, and never express more than one belief of the subject. (We might think of expression here as direct expression: a given statement might indirectly express more than one belief, and indirect expression may allow variation in truth-value.)

When statements do not express a belief (e.g. because the subject is not confident in the truth of what is asserted), I will take it that they express a propositional attitude of a more general sort, one that we might call a *thought*. To think that P, as I will use the term here, is to entertain the hypothesis that P, and a thought that P is an entertaining of the hypothesis that P. Occurrently believing that P entails thinking that P, in much the same way that knowing that P entails believing that P. Like beliefs, thoughts can be true or false. A statement always expresses exactly one thought, and the thought and the statement will always have the same truth-value.

I will take it that there is a relation of *continuance* (or *persistence*) that can hold between thoughts of a subject at different times, and that continuance guarantees that two thoughts that stand in this relation have the same truth-value. I will take it that the thoughts (and derivatively, the statements) of a given subject can stand in a relation of *negation* to each other, and in a relation of *conjunction* (and other logical relations) to each other. And I will take it that a thought can come to be *accepted*, yielding a belief: in this case we can say that the belief is an acceptance of the thought.

We can say that a belief is a priori justified when it is justified independently of experience. A belief is a priori justifiable when it can be justified independent of experience, yielding a priori knowledge. More generally, a thought is a priori justifiable - or more simply, a priori - when it is possible for an acceptance of it to be justified independently of experience, yielding a priori knowledge. Here, apriority abstracts away from contingent cognitive limitations. If there is any possible mental life that starts from a thought and leads to an a priori justified acceptance of that thought, the thought is a priori. A statement is a priori when it expresses an a priori thought. A thought is *epistemically possible* when its negation is not a priori. A statement is epistemically possible when it expresses an epistemically possible thought. One thought (or statement) *implies* another when a conjunction of the first with a negation of the second is epistemically impossible.

Note that the uses of "epistemically possible" in the previous paragraph correspond to deep epistemic possibility, in the idealized sense. In the following discussion, I will be setting aside issues concerning strict epistemic possibility, and non-idealized notions of deep epistemic possibility. So from now on, unqualified talk of epistemic possibility should always be taken as talk of deep epistemic possibility in this sense.

3 Principles of epistemic space

On the picture developed at the start, items of knowledge and belief divide epistemic space. Any item of knowledge or belief excludes some scenarios, and endorses others. More generally, any thought can be held to excludes some scenarios and to endorse others. We can put this by saying that for any given thought (or belief, or item of knowledge), there is a class of scenarios that *verify* the thought and a class of scenarios that *falsify* the thought. Intuitively, a scenario verifies a thought when it corresponds to a way the world might be that the thought endorses, and a scenario falsifies a thought when it corresponds to a way the world might be that the thought excludes.

(Note that verification and falsification as defined here are not evidential notions. It may be epistemically possible, for example, that there is life in black holes for which we can have no evidence. Intuitively, this epistemic possibility is backed by various scenarios in which life exists unperceived in black holes. My thought that there is life in black holes will be verified by these scenarios, not because these scenarios could provide evidence for the truth of the thought, but because they correspond to ways the world might be that the thought endorses. In effect, verification and falsification are broadly representational notions. If one holds independently that truth is tied constitutively to evidence, as on some anti-realist views, it may be that verification will also be tied to evidence; but this tie is not compulsory.)

If we allow that thoughts can have multiple truth-values (such as true, false, and indeterminate), we should also allow that the verification relation can yield the same truth-values. We might say that there is a function that maps an arbitrary scenario S and thought T to a truth-value $verifies(S,T)$. Then S verifies T iff $verifies(S,T)$ is true, and S falsifies T iff $verifies(S,T)$ is false. When S verifies T , we can say that T is true at S . When S falsifies T , we can say that T is false at S . In other cases, it may be that T is indeterminate at S , or perhaps that T has some other truth-value at S .

It is an important part of the picture I am developing that for any subject at a time, there is exactly one way the world *is*, relative to that subject and time. More specifically, relative to any thought, there is exactly one way the world is, relative to that thought. (Recall that thoughts are tokens, embedded in a world.) We can put this by saying that exactly one scenario is *actualized* at the thought. The thought will be true if it is verified by its actualized scenario, and false if it is falsified by its actualized scenario.

So we are postulating a space of scenarios, a function of verification from thoughts and scenarios to truth-values, and a relation of actualization between scenarios and thoughts. These must satisfy at least the following *principles of epistemic space*:

Plenitude: A thought T is epistemically possible iff there exists a scenario S such that S verifies T .

Actuality: For every thought T , there is a unique scenario S such that S is actualized at T .

Truth: For all S and T , if S is actualized at T , then T 's truth-value is $verifies(S,T)$.

Compositionality: For all S and T , if T is a logical composition $\text{comp}(T_i)$, where comp is a truth-function, then $\text{verifies}(T, S) = \text{comp}(\text{verifies}(T_i, S))$.

Parsimony: If scenarios S_1 and S_2 are such that for all possible thoughts T , $\text{verifies}(S_1, T) = \text{verifies}(S_2, T)$, then S_1 and S_2 .

For any thought, there will be an associated set of scenarios that verify the thought. More generally, for any thought, there will be an associated function that maps a scenario to the truth-value of thought at the scenario. We can call this function the thought's *epistemic intension*. A thought's epistemic intension might be thought of as representing its *epistemic content* - the way that it divides epistemic space. Strictly speaking, we might think of an epistemic intension as defined here - relative to an idealized notion of epistemic possibility - as a thought's *ideal epistemic content*, since any two thoughts that are a priori equivalent will have the same epistemic content. (The latter usage would leave open the possibility of a less idealized notion of epistemic content such that thoughts that are nontrivially a priori equivalent can have distinct epistemic content.) We can define the epistemic intension of an statement in a similar way.

One can also postulate similar principles governing the relationship between scenarios and statements. We can say that a scenario verifies an statement if it verifies the thought expressed by that statement, and that scenario is actualized at an statement iff it is actualized at the thought expressed by the statement. Then analogs of Epistemic Plenitude, Actuality, and Truth for statements follow automatically. An analog of Compositionality will follow as long as we allow that statements have compositional relations that mirror the compositional relations of the thoughts expressed. An analog of Parsimony will follow as long as we allow that every possible thought could be expressed by a possible statement.

Of these, Plenitude is the core principle that captures the general picture set out at the beginning. Actualization and Truth capture the special role of the way the world actually is (relative to a thought), among all the ways the world might be. Compositionality ensures that negations and conjunctions, and other logical compositions of thoughts have the truth-values that one would expect. Parsimony says, in effect, that epistemic space contains no redundancy, so that it is a minimal space that accomplishes its epistemic purpose. Of these, it is arguable that Parsimony is not absolutely essential, but the other four principles seem crucial to the notion of epistemic space.

Of course it would be easy to simply postulate that there is a space of scenarios and relations of verification and actualization that satisfy these principles. But this would raise the question of whether the postulation is coherent. We can make a positive case for coherence by *constructing* epistemic space. That is, we can identify the space of scenarios with a space of abstract objects that we have prior reason to believe is coherent, and we can make the case that there are relations of verification and actualization over this space that satisfy the relevant principles. In the following sections, I explore two ways of making such a construction.

There is a further principle worth mentioning:

Maximal Thoughts: For every scenario S and every subject, there exists a possible thought T such that for all possible thoughts T' of the subject, S verifies T' iff T implies T' .

We might think of T as the thought that S is actual. Where S is a maximally specific epistemic possibility, T will be a maximally specific coherent thought. Intuitively, S will verify T' iff the thought that S is actual implies the thought that T' (that is, if the thought " S is actual and $\sim T'$ " is epistemically impossible), and S will falsify T' iff the thought that S is actual implies the thought that $\sim T'$ (that is, if the thought that " S is actual and T' " is epistemically impossible).

Let us say that a *maximal thought* is an epistemically possible thought T such that there is no thought T' such that T' implies T while T does not imply T' . (Intuitively, if both conjunctions were coherent, then T could not be a maximally specific coherent thought.) It is easy to see that if Maximal Thoughts is true, then for any S , the corresponding thought T is a maximal thought in this sense. (If it were not, then S would verify both T_1 and $\sim T_1$ for some T_1 , so it would verify the epistemically impossible $T_1 \& \sim T_1$.) We can call T a maximal thought corresponding to S .

The Maximal Thought principle is not essential to the notion of epistemic space. It is not entirely obvious that maximal thoughts can exist, and the idea of epistemic space may still be coherent even if there is a problem with maximal thoughts. But if maximal thoughts can exist, the principle certainly provides a useful aid in the analysis of epistemic space.

4 Scenarios as centered worlds

The most natural way to think of scenarios, at least initially, is as possible worlds. In a way this is trivial - scenarios are defined as possible (in some sense) ways a world might be (in some sense). But the notion of possibility invoked here differs from the notion of possibility that is usually associated with possible worlds: it is a sort of epistemic possibility, whereas possible worlds are usually understood to be associated with a sort of "metaphysical" possibility. Still, the question arises as to whether possible worlds understood in the latter sense might serve to help us model the space of scenarios, at least indirectly. That is: can we use the space of metaphysically possible worlds to construct a space of scenarios, and to make the case for a verification relation between scenarios (so understood) and thoughts?

I think we might. The intuitive idea is simple: to every possible world W , there corresponds a very specific (deep) epistemic possibility: the epistemic possibility that W is actual. Given any specific epistemic possibility of this sort, it will be epistemically compatible with some thoughts, and incompatible with others. One might then proceed as follows: (i) identify the space of scenarios with the space of possible worlds; (ii) say that a world W verifies a thought T when the hypothesis that W is actual is epistemically compatible with T ; and (iii) say that a world W is actualized at a thought T when W is the world containing T .

The idea here is intuitive, since I have invoked the intuitive idea of a hypothesis, and of a hypothesis

being compatible with a thought. One could give more flesh to the idea by invoking the *thought* that W is actual instead, and holding that W verifies T iff the thought that W is actual is epistemically compatible with T (where two thoughts are epistemically compatible when their conjunction is epistemically possible). As with the maximal thought principle, it is not obvious that such specific thoughts can exist for all W, so we do not want to rest the framework on their existence. And even if we take this approach, there is still an intuitive element, in that we have appealed to the intuitive idea of a thought that W is actual. But in any case, this way of thinking about things at least gives a sense of why one might try to model scenarios with possible worlds.

I think that this sort of approach is on the right track. But there are a few obstacles to understanding scenarios in terms of possible worlds, some of which we have already seen. The obstacles fall into four classes, which we can class under the four headings: indexicality, rigidity, strong necessity, and parsimony.

(1) Indexicality

The first obstacle is a problem for any view that identifies scenarios with objective states of affairs. As a number of philosophers have argued, there are certain questions that may be left open by any amount of objective information about the world. One can put this as follows. Let D be a full objective specification about the world. Let T be an indexical claim, such as "I am a philosopher" or "It is raining here now" or "Today is Friday". Then in each case, it may be that both D&T and D&~T is epistemically possible. That is, the information in D may not enable one to settle the status of T. To settle these indexical claims, one needs to be able to *locate* oneself and the current location and time with respect to an objective description of the world; and this locating information cannot be derived from objective information. So objective information is not maximal information.

The problem this poses is clear. In these cases, D&T and D&~T are epistemically possible, so both are verified by a scenario; and they are epistemically incompatible, so no scenario verifies both. But if D is a complete objective description of the world, then there will plausibly be only one world (objectively understood) in which D is the case. And even if there is more than one objectively indistinguishable D-world, it seems that there is no room in such a space of objective worlds for a division between worlds in which today is Saturday and worlds in which it is not.

The basic problem is in satisfying Plenitude. Maximal epistemic possibilities are differentiated not just by objective information but also by indexical information, and there are not enough possible worlds to go around. To model epistemic space, we need a more fine-grained space that allows indexical differentiation.

This suggests a natural solution: we need to *supplement* possible worlds with some further indexical structure. Specifically, we can identify scenarios with *centered worlds*: worlds marked with a "center" consisting of a subject and a time within the world. (Equivalently, centered worlds might be identified with ordered triples of a world, a subject in that world, and a time in that world.) We can think of the

center as a sort of "you are here" marker, corresponding to a hypothesis about the subject's identity and location. If W' is a centered world, the hypothesis that W' is actual will be a sort of indexical hypothesis about the world and the subject's location within it. Putting things in terms of thoughts: if W' consists of an ordinary world W centered on a subject and a time, my thought that W' is actual will be the thought that I am now at the center of W' (i.e. that W is actual, I am the subject indicated, and now is the time indicated).

When scenarios are modeled by centered worlds, the problems above can be handled straightforwardly. My thought that I am a philosopher will be verified by all centered worlds in which the subject at the center is a philosopher. My thought that it is raining here now will be verified by all centered worlds in which it is raining at the location of the marked subject at the time marked at the center. My thoughts T1 and T2 above will be verified by different classes of centered worlds, with an objective world in common but different locations for the center. And so on. Intuitively, this is just how we would expect these thoughts to divide the space of epistemic possibilities.

It is useful to stipulate that the marking of centered elements in a centered world is optional. This way, we can accommodate the (arguable but plausible) a posteriority of claims such as "Subjects exist" and "The universe is temporal". If we allow centered worlds without marked subjects or times, then there will be subjectless scenarios and timeless scenarios to falsify these claims. There can even be an empty scenario to verify "Nothing exists", which is arguably a deep epistemic possibility. (Here I assume that "I exist" is a posteriori, being justified by experience. If someone holds that "I exist" is a priori, then they can require that centered worlds contain marked subjects.) It may also be that we sometimes need additional optional marked information at the center of a world. For example, for the verification of certain demonstrative thoughts (e.g. Austin's "Two Tubes" puzzle - "that spot is red", in a symmetrical visual field with two spots), one needs to allow one or more marked experiences at the center of a world, to distinguish otherwise indistinguishable contents. (For more discussion, see...) But the general framework is much the same.

(2) Rigidity

The second problem has already been mentioned. There are subjects for whom it is epistemically possible that Hesperus is not Phosphorus: more specifically, there are subjects for whom the statement "Hesperus is not Phosphorus" expresses an epistemically possible thought. But on the usual understanding of possible worlds, "Hesperus" and "Phosphorus" are *rigid designators*: they pick out the same object (here the planet Venus) in all possible worlds. If so, then there is no possible world satisfying "Hesperus is Phosphorus". Something similar applies to the epistemic possibility that water is not H₂O ("water" picks out H₂O in all worlds, so no world satisfies "water is not H₂O") and the epistemic possibility that my greatgrandparents were or were not cousins (I have the same ancestors in all worlds, so if my greatgrandparents are not cousins, no world satisfies "my greatgrandparents are cousins"). Adding centers to the possible worlds does not help with this. So it may seem that if scenarios are centered worlds, then Plenitude cannot hold: there are cases in which an statement (e.g. "water is not H₂O") is epistemically possible, but there is no centered world satisfying the statement.

This conclusion would be premature, however. There will be a problem for Plenitude only on the assumption that a possible world *verifies* an statement precisely when it *satisfies* that statement. And this points immediately to a possible solution: we can hold that verification and satisfaction are quite different relations. On such a view, it may be that for epistemically possible statements such as the above (e.g. "water is not H₂O"), then although no world satisfies the statement, there are nevertheless worlds that verify the statement.

In fact we have already seen reason to accept that verification and satisfaction are different relations. Above, we saw that "I am a philosopher" is verified by any centered world in which the being at the center is a philosopher. But it is false that any such world satisfies "I am a philosopher". For example, consider a counterfactual world in which David Chalmers is a mathematician and George Bush is a philosopher, and consider a centered version on this world, centered on George Bush. Then by the standard sort of evaluation, this is a world in which I am not a philosopher (instead I am a mathematician) - that is, the world satisfies "I am not a philosopher". But nevertheless, the world verifies "I am a philosopher".

The distinction is also suggested by the intuitive test for verification of a thought T given above: is the hypothesis that W is actual epistemically compatible with T? Take a "Twin Earth" world W that is superficially like our world, and in which the oceans and lakes around the center are filled with (clear, drinkable) XYZ. It is epistemically possible (i.e. not ruled out a priori) that W is actual. The epistemic possibility that W is actual is epistemically compatible with the thought that water is XYZ, and is epistemically incompatible with the thought that water is H₂O. If so, then W verifies "water is not H₂O". At the same time, Kripke and Putnam have argued that W is a counterfactual world in which XYZ is not water, and in which water is still H₂O (if it exists). So W does not satisfy "water is H₂O", even though W verifies "water is H₂O".

What is going on here? The root of the difference is that satisfaction is a broadly subjunctive notion, concerning ways the world might have been. To determine whether the Twin Earth world satisfies "water is XYZ", we can ask: if the liquid in the oceans and lakes *had been* XYZ, would water *have been* XYZ? If Kripke and Putnam are correct, the answer is no. Verification, on the other hand, is a broadly epistemic notion, concerning ways the world might be. To determine whether the Twin Earth world verifies "water is XYZ", we can ask: if the liquid in the oceans and lakes *is* XYZ, *is* water XYZ? This indicative conditional behaves epistemically, turning on whether there is an appropriate epistemic relation between the antecedent and the conclusion. And in this case, the epistemic relation is present, so the answer is yes.

The notion of possibility that is operative in most contemporary discussions is what one might call *subjunctive possibility*: P is subjunctively possible if it might have been that P. (Kripke is quite explicit about this, for example.) The notion of possibility that is operative in the epistemic realm is epistemic possibility: P is epistemically possible if it might be that P. It is already a familiar point that these two notions of possibility behave differently. It is not surprising that there are correspondingly different relations between thoughts, statements, and possible worlds.

One can use verification and satisfaction to define parallel notions in the epistemic and subjunctive realms. For example, we defined the epistemic intension of an statement A or a thought T as the function mapping a scenario S to verifies(S,A) or verifies(S,T). If scenarios are centered worlds, then the epistemic intension is a function from centered worlds to truth-values. We can similarly define the *subjunctive intension* of A or T as the function mapping a possible world W to satisfies(W,A) or satisfies(W,T) (where we hold that a world satisfies a thought if it satisfies an statement that expresses that thought). If scenarios are centered world, these intensions will have closely related domains (the same up to centering), but for a given thought or statement, the intensions themselves will be very different.

In a similar way, one can postulate both epistemic intensions and subjunctive intensions for concepts and for tokens of expressions smaller than sentences, mapping. When an expression is a rigid designator - such as "I", "water", or "Aristotle" - its epistemic intension and subjunctive intension will usually behave quite differently. The subjunctive intension of a rigid designator will pick out the same object in all worlds, but its epistemic intension will not. For example, the subjunctive intension of my concept "I" will pick out David Chalmers in all worlds (mirroring the subjunctive necessity of "I am David Chalmers" - that is, I could not have failed to be David Chalmers). But the epistemic intension of "I" picks out whichever being is marked at the center of a world, irrespective of whether he is David Chalmers. Similarly, the subjunctive intension of "water" picks out H₂O in all worlds (if Kripke and Putnam are correct), but its epistemic intension picks out different substances in different worlds, as we have seen.

The same applies to (tokens of) names such as "Aristotle". The subjunctive intension of a name picks out the same person across worlds, but the epistemic intension cannot. To see this, note that it is (deeply) epistemically possible for me that Plato was Aristotle: this is unlikely, but not ruled out a priori. So there must be a scenario verifying my statement "Plato is Aristotle"; and if scenarios are centered worlds, there must be a centered world verifying "Plato is Aristotle". One such is a world W in which an individual lived a double life, publishing one set of (familiar-looking) books as "Plato" and another as "Aristotle", and in which the community at the center has no knowledge of that distinctness. Intuitively, the hypothesis that W is actual is (deeply) epistemically possible, and it epistemically entails the hypothesis that Plato is Aristotle. If Kripke and Putnam are right, this is not a world *satisfying* my statement "Plato is Aristotle", but it may nevertheless be a world *verifying* "Plato is Aristotle".

This suggests a broadly *two-dimensional* approach to meaning and possibility, and as such may recall the two-dimensional approaches developed by Kaplan, Stalnaker, Evans, and others. I have explored the similarities and differences elsewhere. One important difference, however, is that an epistemic intension is not defined in terms of the *context-dependence* of a thought or statement's truth-value. The epistemic intension is a matter of *epistemic dependence*, turning on whether a thought T is epistemically compatible with the hypothesis that W is actual. Nothing here gives any special role for a copy of the thought T at the center of W, and the epistemic intension will usually be defined over worlds in which any such token is absent. The epistemic intension will even be defined over worlds in which language and thought are absent altogether. Witness "language does not exist", or "no-one exists", each of which is arguably deeply epistemically possible, and which will be correspondingly verified by languageless worlds and by subjectless worlds respectively.

Note that I have not given a formal definition of what it is for a centered world to verify a thought, but only an intuitive account: W verifies T when the hypothesis that W is actual is epistemically compatible with T. A formal definition would require two things. First, we would have to rigorize talk of the epistemic compatibility of hypothesis and thoughts, either by formally defining and characterizing epistemic possibilities, or by dispensing with hypotheses and appealing to the thought that W is actual instead. Second, we would have to give an account of what is involved in the hypothesis that W is actual. One natural way to do this would be to distinguish a sort of canonical way of describing possible worlds, so that for any centered world W there will be a canonical description D (including indexical claims to capture the centering). Then the hypothesis or thought that W is actual might simply be the hypothesis or thought that D.

So one possible formal definition would be the following: a centered world W verifies a thought T when T is epistemically compatible with a thought expressed by D, where D is a canonical description of W. Of course this definition requires an account of canonical descriptions, and it requires that the relevant maximal thoughts can exist. I discuss the details of this and other formal definitions elsewhere. For present purposes, it is sufficient to note that the general approach to verification is *prima facie* coherent, and that the phenomena of rigidity pose no fatal problems.

(3) Strong necessities

We have seen that the existence of a posteriori necessities such as "water is H₂O" poses no deep problem for the picture of scenarios as centered worlds, as long as we distinguish verification from satisfaction. It is not hard to see that all of the a posteriori necessities suggested by Kripke can be handled in a similar way: in these cases, when N is an a posteriori necessity there is plausibly a centered world verifying N. But on some philosophical views, there exist a posteriori necessary truths that do not have this property. We can call these *strong necessities*.

Consider a theist view on which it is necessary that an omniscient being exists (e.g. because it is necessary that God exists and is omniscient), but on which it is not a priori that an omniscient being exists (e.g. because it is not a priori that God exists). On such a view, "There is an omniscient being" is necessary but a posteriori, and "There is no omniscient being" is (deeply) epistemically possible. On such a view, it will not merely be the case that no possible world satisfies "There is no omniscient being". It seems clear that no world verifies "There is no omniscient being". That is, all possible worlds W are such that the hypothesis that W is actual is epistemically incompatible with the thought that an omniscient being exists. There are no relevant two-dimensional phenomena involving rigidity here: it simply seems that there are not enough worlds to go around. If this view is correct, then a model on which scenarios are centered worlds is incompatible with Plenitude.

There are some other views on which the same applies:

(i) A particular strong "strong laws" view on which the fundamental laws and properties instantiated in

our world are the fundamental laws and properties of every possible world. Let say the view also holds (plausibly) that fundamental laws are a posteriori. On this view, a denial of the law of gravity (say) will be deeply epistemically possible, but there will be no possible world satisfying this denial, and there will also be no possible world verifying the denial.

(ii) A materialist view on which truths Q about experience are necessitated by the conjunction P of physical truths, but on which Q is not a priori derivable from P. Here, a psychophysical conditional "P and not-Q" will be epistemically possible. It is not hard to show that if there is even a possible world verifying this conditional (as in the Kripkean cases), problems for materialism ensue. So some materialists deny that even a verifying world exists. If so, the conditional "If P, then Q" is a strong necessity.

(iii) A view on which there are mathematical claims M - perhaps the Continuum Hypothesis? - that are true and are necessary, but are not knowable a priori by any possible being. On such a view, it seems that M will be a strong necessity: $\sim M$ will be epistemically possible, but verified by no possible world.

Other such views could be developed: e.g. one on which moral claims can be true and necessitated by natural truths, without being a priori derivable from natural truths; or a similar view about vague claims. In each of these cases, the distinction between verification and satisfaction does not seem to help. If the views in question are correct, there are simply not enough possible worlds to verify all epistemically possible thoughts and statements.

The simplest response to this problem, and the response that I believe is correct, is to deny that strong necessities exist. In each case, the claim in question about necessity is at least controversial. In some cases, proponents claim support from the Kripkean cases, but these cases give no reason to believe in this much stronger phenomenon. In fact, one can argue in reverse: the fact that the link between epistemic possibility and verification by possible worlds is so strong elsewhere gives reason to believe that these claims are incorrect. One can also argue that there are deeper problems with these views. But I have argued for these claims elsewhere, and will not repeat those arguments here.

It is at least clear that these views provide no *clear* reason to reject the model of scenarios as centered worlds, since in no case is the view in question clearly true. Still, the existence of these views entails that the claim that scenarios can be modeled by centered worlds will be at least as controversial as the denial of the views. And it would be desirable to give an account of scenarios that even holders of these views could accept. If so, that provides at least some reason to look at other models of scenarios.

(4) Parsimony

So far, we have looked at problems of the sort: there are *not enough* possible worlds to act as scenarios. But there are also potential problems of the sort: there are *too many* possible worlds to act as scenarios. That is, while the problems above are mostly problems for Plenitude, one can also raise problems for Parsimony. In particular, it seems that there exist groups of centered worlds such that any possible

thought is equally verified or falsified by any world in the group. If so, it seems that each world in the group corresponds to the same scenario.

One way this can happen is with symmetrical worlds. Say that a world is mirror-symmetrical, and consider centered worlds W_1 and W_2 centered on corresponding subjects on each side, at the same time. Then it seems that there is no thought T such that the thought that I am in W_1 is epistemically compatible with T , while the thought that I am in W_2 is epistemically incompatible with T . The same goes for a world with a cyclic Nietzschean eternal recurrence of indistinguishable cycles, extending indefinitely into the past and the future. If we take a group of centered worlds W_i centered on corresponding subjects and times in different cycles, then it seems that for any T , if one world W_i verifies T , then all worlds W_i verify T . In these cases, it seems that the different centered worlds all correspond to the same epistemic possibility, violating Parsimony.

Parsimony might also be violated if possible worlds can contain inconceivable features. Say that there are two possible worlds W_1 and W_2 that are otherwise indistinguishable, except that at a certain point they contain different features F_1 and F_2 . And say that F_1 and F_2 are inconceivable, in the sense that there is no possible concept that refers to F_1 , or to F_2 . Then it seems that there will be no possible thought T such that T is verified by W_1 but not W_2 .

Finally, suppose that (as some believe) there are qualitatively indistinguishable possible worlds. Take two identical twins Bill and Bob in the actual world. Some argue that there can be qualitatively indistinguishable worlds W_1 and W_2 such that only Bill exists in W_1 and only Bob exists in W_2 . Then it is plausible that W_1 and W_2 do not correspond to distinct epistemic possibilities: any thought verified by W_1 is also verified by W_2 . If so, Parsimony is violated.

Of these three cases, the last two are controversial and might be denied. But the first is relatively uncontroversial, and the other two at least raise problems. So it seems that the space of centered worlds and the verification relation, as we understand them, do not satisfy Parsimony.

One could respond in different ways. One might simply jettison Parsimony, holding that it is an inessential principle: certainly it seems less essential than Plenitude. One might also modify the picture slightly, by identifying scenarios with *equivalence classes* of centered worlds, where the worlds in groups such as the above will all fall into the same equivalence class. Either response will still allow a serviceable construction. Still, both responses suggest that there is at least a mild mismatch between scenarios and centered possible worlds.

What is the upshot of the four obstacles to identifying scenarios with possible worlds that we have discussed? The obstacles due to indexicality and rigidity can be overcome relatively easily, by invoking centered worlds and distinguishing verification from satisfaction. The obstacle due to strong necessity can be denied, and the obstacle due to parsimony can be dealt with as above.

Still, the last two obstacles suggest that while centered worlds may do a good job of modeling scenarios,

the match is not perfect. The existence of philosophical views on which there are strong necessities suggests that even if these views are misguided, an analysis of scenarios as centered worlds will be at least mildly controversial. Because it makes a substantive (if plausible) claim about the relationship between possible worlds and epistemic possibility, it goes beyond a surface analysis of epistemic possibility itself. The problems with parsimony also suggest a slight conceptual mismatch between the notions. So while centered worlds may provide a very useful way of thinking about scenarios, it may also be useful to look at other notions.

5 The epistemic construction of scenarios

The obstacles in the previous section all have a common source. They arise because we are taking a class of entities - the possible worlds -- developed in the service of a *different* notion of possibility (what might have been the case), and adapting it to help analyze the notion of epistemic possibility (what might be the case). It is inevitable that this adaptation will lead to certain complications. An alternative strategy suggests itself. Instead of adapting a different modal space, we might construct the space of scenarios directly, by a construction grounded in epistemic notions. In particular, we might take (deep) epistemic possibility as basic, and proceed from there. In this way, we can give an account of epistemic space in its own right.

The natural way to proceed is to identify scenarios with constructions out of thoughts or statements. We already have a notion of epistemic possibility that applies to these entities, and this notion can be exploited to construct scenarios directly. Of course we will need to appeal to *possible* thoughts and statements, since it is unlikely that all the thoughts and statements we need will exist in the actual world. For present purposes, I will understand these as subjunctively possible thoughts and statements - thoughts and statements that might have existed. Because of this, our construction of epistemic space will not be entirely independent of subjunctive modal notions, but the relationship will be much more indirect than in the previous constructions, so that the resulting epistemic space may be quite different from subjunctive modal space. An alternative strategy would be to appeal to the *epistemic* possibility that certain thoughts and statements exist, and bootstrap the construction from there, but I will not pursue that strategy here. Henceforth (to avoid clutter and also to avoid confusion with the existing and distinct notion of an epistemically possible thought), reference to thoughts and statements should be understood as reference to possible thoughts and statements.

I will work here with thoughts rather than statements, but everything here could straightforwardly adapted to statements. The approaches pursued here will start with the possible thoughts of a subject at a time, and so will first define a space of scenarios relative to a given subject and time. After that, we can look at defining a common space across subjects and times.

The simplest strategy is to identify scenarios with equivalence classes of thoughts. In particular, we can appeal to the class of *maximal thoughts*. As defined previously, a maximal thought is an epistemically possible thought T such that there is no epistemically possible thought T' such that T' implies T while T does not imply T' . We can say that two maximal thoughts T_1 and T_2 are *equivalent* if T_1 implies T_2 and

T2 implies T1. Then we can say that a scenario is an equivalence class (under this equivalence relation) of maximal thoughts, and that a scenario verifies a thought T if a maximal thought in its equivalence class implies T.

It is easy to see that this construction will satisfy Compositionality: this follows from the analogous principle about implication. The construction will also satisfy Parsimony: if two maximal thoughts imply the same thoughts, then they will imply each other, and will be members of the same equivalence class.

In order to satisfy Plenitude, we need the following principle: for every epistemically possible thought T, there is a maximal thought that verifies T. It is not absolutely obvious that this principle is correct: it is not absolutely obvious that maximal thoughts exist (perhaps for every possible thought, there will be a more specific thought?), and it is not absolutely obvious that there will be one for every epistemic possibility. Still, there is no obvious problem with the idea. And it is at least intuitively reasonable to suppose that from a given thought, a maximal thought could be obtained as a sort of maximal conjunction, by conjoining other thoughts or their negations until the limits of epistemic possibility are reached.

It is not obvious how to define the actualization of a thought: the maximal thought corresponding to how things are from the point of view of a given subject at a given time, thinking a given thought. It is tempting to say: a maximal thought M is actualized at a thought T if, were the subject to think M at the same time as T, M would be true. But this cannot work: any M that implies "I am not thinking a maximal thought" will fail this test, but intuitively, an M like this is required to capture how things are for most subjects. In effect, we want M to correctly describe how things were *before* M was thought. I will not pursue the project of defining actualization here. Instead I will take it that we have a reasonably good intuitive grip on the notion, and I will simply assume a relation of actualization between maximal thoughts and thoughts, such that a thought is true iff it is implied by its actualization.

If one has worries about maximal thoughts, one might also construct scenarios via maximal *classes* of thoughts. Let us say that a class of thoughts is *epistemically possible* when, intuitively, the conjunction of thoughts in the class would be epistemically possible. Let us say that a *complex* is an epistemically possible class of thoughts. We can say that a complex C implies a thought T when the union of C with $\{\sim T\}$ is not epistemically possible. A complex C implies a complex D when C implies every thought in D. We can say that a maximal complex is a complex C such that there is no complex C1 such that C1 implies C while C does not imply C1. We can then identify a scenario with equivalence classes of complexes under mutual implication.

The definition of a complex appeals to the notion of an epistemically possible class of thoughts, which in turn appeals to the notion of the epistemic possibility of an arbitrary conjunction of thoughts. If we allow that for every class of thoughts (that is, of possible thoughts of a subject at a time), there is some possible thought that is their conjunction, then we can define this notion wholly in terms of the epistemic possibility of thoughts. But even if we deny this, perhaps because of in-principle cognitive limitations, it is arguable that the notion of an epistemically possible class of thoughts still makes sense. One could try to define it in various ways (e.g. via the claim that for every thought T that is a conjunction of thoughts in

the class, T is epistemically possible), or one could simply assume the notion. In any case, the claim that every thought is implied by some complex is weaker than the claim that every thought is implied by some maximal thought, so it is useful to have this notion on the table.

In any case, given the assumptions above, we can define the space of scenarios (relative to a subject) out of the possible thoughts of that subject, and the five central principles will be satisfied.

A question arises: will the space of scenarios (so defined) have interesting structure? Or will it have the structure simply of arbitrary classes of thoughts? This question depends on the extent to which thoughts are epistemically related, and to which there is epistemic entailment between different classes of thoughts. The question is too large to say much about here, but we can at least say a little.

Earlier we noted that at least some thoughts (perhaps all) are composed of concepts. Let us say that a thought T is *composed from* a class C of concepts if T is composed only of concepts in C. Then we can say that a class C of concepts is a *basis class* if all thoughts are implied by some thought composed from C. (Note that we do not require that all thoughts are *equivalent* to some thought composed from C.)

It is plausible that basis classes exist. The class of all concepts is plausibly a basis class (unless there are thoughts that are neither composed of concepts nor verified by any thoughts that are composed of concepts). Given any basis class C, then given any maximal thought M, there will be a thought M' composed from C such that M' implies M. Then M implies M' (by maximality of M), and M and M' are equivalent maximal thoughts. It follows that for any basis class C, we can construct arbitrary scenarios using only concepts in C.

The question then arises: how small can a basis class of concepts be? This is one of the deepest questions in philosophy. But here, I will simply note that it is plausible that many concepts are eliminable from a basic class. I have argued elsewhere that for many thoughts (e.g. "water is H₂O", "Oswald killed Kennedy") there are statements about the world that are *nontrivially sufficient* to settle the truth of the thought. The exact form of these statements is not important here (plausible candidates include qualitative descriptions of the world in physical, mental, and indexical terms). What matters is there exist such nontrivially sufficient statements: sufficient in that they imply the thought or its negation, and nontrivially sufficient in that the statements in question do not invoke the main concepts from the original thought (e.g. "water") or their cognates. This phenomenon is not restricted to the actual world: in general, for an arbitrary epistemic possibility, there are statements about that possibility that are nontrivially sufficient to settle the status of the thought with respect to that possibility. If so, it follows that the main concepts in these thoughts are eliminable from a basis set of concepts.

Just how far this can be taken is an open question. I think it is plausible that a basis set can be stripped down quite a long way. I have argued that at least for possibilities in the vicinity of the actual world, a basis set consisting of microphysical concepts, phenomenal concepts, and indexical concepts (plus logical and mathematical concepts and other "framework" concepts) suffices. One might argue that microphysical concepts are themselves eliminable in favor of general spatial, temporal, causal, and

experiential concepts, and perhaps ultimately in favor of phenomenal concepts plus the concept of causation. But I will not defend any such strong claim here. (Note again that the claim is not quite as strong as it may sound, as we do not require that every thought be *equivalent* to a basis thought.) What matters is that at least some concepts are eliminable from a basis set. To the extent that there are eliminable concepts, there will be nontrivial structure in the relation between scenarios and thoughts.

Having defined a space of scenarios for each subject at a time, we must now ask: how can we define a common space of scenarios for all subjects? (For ease of discussion, I will talk simply of subjects instead of subjects-at-times.) This requires some principle of *translation* between the maximal thoughts (or the complexes) of one subject, and those of another. We could simply stipulate that there is a relationship of translation among maximal thoughts, such that for any two subjects, a maximal thought of one subject is translated by exactly one maximal thought of another (at least up to a priori equivalence). This would not be very informative, however. Instead, it is useful to ground translation in some more basic notion of translatability.

To translate maximal thoughts, it suffices to define a translation relation across basis classes of concepts in all subjects. In more detail, let us say that a *basic translation relation* is an equivalence relation (translation) among a class C of *translatable concepts* across all subjects - yielding a derivative class of translatable thoughts composed from C, with a translation relation among them - such that:

- (i) For any two subjects, any translatable concept of the first is translated by some concept of the second.
- (ii) If a thought T1 translates a thought T2, then T1 is epistemically possible iff T2 is epistemically possible.
- (iii) For any subject, the translatable concepts of that subject are a basis class.

It is easy to see that any basic translation relation yields a mapping between scenarios across subjects. Given scenarios S1 and S2 of subjects, we can stipulate that S1 translates S2 iff for any pair of translatable thoughts of the two respective subjects, S1 verifies T1 iff S2 verifies T2. Then for any two subjects and any scenarios S1 of the first subject, there will be a unique S2 that translates S1.

(To see this, note first as a lemma that if T1' implies T1 and if these two thoughts are respectively translated by T2' and T2, then T2' implies T2: the epistemically impossible T1' & ~T1 will be translated by T2' & ~T2, which will be epistemically impossible by (ii). Now take any two subjects and a scenario S1 of the first subject. By (iv), there will be a translatable maximal thought T1 associated with S1. By (i), this will be translatable by a thought T2 of the second subject. T2 is maximal: if it were not, it would be implied by a translatable maximal thought it does not imply, so (by the lemma) T1 would be implied by a (translated) thought it does not imply, so T1 would not be maximal. Given two any two translatable thoughts T1' and T2' of the subjects, the lemma entails that T1 implies T1' iff T2 implies T2', so S1 implies T1' iff S2 implies T2'. So S2 translates S1. If any scenario S3 translates S1, then S3 will verify T2 (since S1 verifies T1 and T2 translates T1), so S3 verifies T2, and S3=S2. So S2 is the unique scenario

that translates S1.)

So for a common space of scenarios, we need not start with a translation relation over all concepts: a translation relation on a class of basis concepts will suffice. That is, we need a notion of *epistemic translatability* that clearly applies to some concepts, and is such that the concepts to which it applies forms a basis class. I will not try to define this relation here, but I will say a few words about it.

For two thoughts to be epistemically translatable, then intuitively, they must divide epistemic space in the same way. We can get an initial grip on the notion from the discussion of centered worlds above. Two thoughts "I am a philosopher" by me and a friend will each be verified by all centered worlds in which the being at the center is a philosopher (at least, if there are no relevant differences in our concepts of "philosopher"); so these thoughts will be epistemically translatable. The same goes, plausibly, for two thoughts "my greatgrandparents are first cousins", and the like. In a similar way, for two concepts to be epistemically translatable, they must at least have the same epistemic intension across centered worlds. It seems that any two "I" concepts will be epistemically translatable, and that two "first cousin" concepts are epistemically translatable, and so on.

Clearly, epistemic translatability is not a notion that requires sameness of *extension*. This is illustrated by the case of "I", and by the more general point that two concepts can have the same epistemic intension across centered worlds but quite different extensions, due to embedding in different centered worlds. The crucial property of epistemic translatability is requirement (ii) above: for this requirement, sameness of extension is irrelevant (and most extension-dependent notions of translatability will not meet it). Rather, what matters is something like sameness of inferential role, and in particular isomorphism of a priori connections.

I will not try to formally define this notion of epistemic translatability here. But even without a formal definition, I think that we have a clear grasp of an epistemic translatability relation over a wide class of concepts. Translatability intuitions become difficult where extension-dependent (or otherwise environment-dependent) concept types are involved, as with the concepts expressed by names and natural kind terms, and also with concepts that are expressed by terms used with semantic deference. But even when these concepts are eliminated, a robust class of concepts remains (such as those above), with a relatively uncontroversial translation relation among them. And it can be plausibly argued that these concepts form a basis class (or something close to it), so that we have a basis translation relation. If so, this yields a common space of scenarios, and a construction of epistemic space.

(There is one obstacle to a basic translation relation here, involving demonstrative concepts in the sort of case mentioned before. Say I demonstrate one of two spots in a symmetrical visual field and think "this is red" (T); and say another possible subject in the same situation thinks two analogous "this is red" thoughts (T1 and T2), one for each spot. Then T plausibly is no better or worse translated by T1 or T2; and T1 or T2 do not translate each other, as they do not imply each other. So T is translated by neither T1 nor T2. More generally, it seems that no possible thought of the other subject is appropriate to translate T. For the same reason, no possible concept of his is appropriate to translate my "this" concept, and there seems to be no underlying translatable basis concept either. I think the moral is that experiential

demonstratives are not strictly translatable, but are *sui generis*.

If this is right, then the space of scenarios for a subject will involve a common translatable basis plus a further basis of untranslatable demonstratives. We can either have a mildly different space of scenarios for each subject (slightly modifying the original picture), or an extended common space with all possible demonstratives (such that some subjects cannot be related to some scenarios). Either way, demonstrative thoughts will never have identical epistemic content across subjects. But we can define an *epistemic isomorphism* relation between thoughts and between scenarios, such that T stands in this relation to both T1 and T2, which will allow good enough cross-subject comparisons for most purposes.)

It is worth noting that insofar as we have an independent grasp on a notion of epistemic translatability of (at least some) thoughts, this imposes a further constraint on epistemic space that might be worth listing alongside the original principles of epistemic space:

Translatability: If thoughts T1 and T2 are epistemically translatable, then for all scenarios, S verifies T1 iff S verifies T2.

To be made precise, this principle would require a precise and independently grounded definition of epistemic translatability across some class of concepts and thoughts. But even in the absence of a precise definition, insofar as we have an intuitive grasp of epistemic translatability (at least in clear cases), the Translatability principle yields a substantial constraint that epistemic space should satisfy.

Instead of further pursuing the strategy based directly on epistemic translatability of thoughts, I will instead pursue a closely related linguistic strategy for constructing scenarios instead. This strategy starts with *sentences*, where these are understood as types within a language rather than as tokens.

It is natural to try to construct scenarios by invoking an epistemic possibility operator over sentences, and proceeding from there. But there is an obstacle: many words and sentences exhibit a sort of *epistemic variation*, so that one token of a sentence may be a priori and another not. This happens most obviously with ambiguous terms, and terms with context-dependent criteria of application. It can also happen with names and natural kind terms. For example, if Leverrier uses "Neptune" as a name for whatever perturbs the orbit of Uranus, and if his wife picks up the term from him without knowing this, then their two corresponding concepts may have different epistemic intensions. To see this, note that "Neptune perturbs the orbit of Uranus, if it exists" expresses an a priori thought for Leverrier, but not for his wife. So the two statements will have different epistemic intensions, arising from difference in the subjects' concepts of "Neptune". So sentences like this are not well-suited to constructing epistemic space.

To deal with this obstacle, we can restrict attention to a special class of *epistemically invariant* expressions. This class should at least have the property that if a sentence S contains only epistemically invariant expressions, then if one token of S is epistemically possible, all tokens of S will be epistemically possible (setting aside tokens of S that are used deferentially.) And more generally, these expressions should be such that (intuitively) their epistemic intensions across centered worlds cannot

differ.

To obtain such a class, we can exclude names and natural kind terms, which allow phenomena such as the above. We can regiment language to eliminate ambiguity. And we can regiment language to eliminate terms with context-dependent criteria of application (especially criteria that vary with speaker's intentions or with conversational context; simple dependence of extension on the environment is allowable). The terms that remain are epistemically invariant. These terms will include pure indexicals such as "I", "now", and "here", and logical and mathematical terms. Also included will be pure descriptive terms: perhaps terms such as "philosopher", "circle", "justice", "object", "cause", "perceive", and "conscious", at least if these are regimented to avoid contextual variation.

Let us say that a language is epistemically invariant if all terms in it are epistemically invariant. English is not epistemically invariant, although it plausibly contains epistemically invariant sublanguages (by restricting the lexicon). But there are certainly many possible languages that are epistemically invariant, and some of them will have significant expressive power.

If L is an epistemically invariant language and S is a sentence in L, then if one token of S is epistemically possible, all tokens of S are epistemically possible (again, setting aside tokens used deferentially). So here we can associate epistemic possibility with types: a sentence S is epistemically possible if any (nondeferential) token of S is epistemically possible. It is reasonable to hold that any sentence of L can be used in principle by an arbitrary subject, expressing a thought. We can define an implication relation between sentences of L and arbitrary thoughts: a sentence S implies a thought T if the thought that the subject would express with S implies T. The epistemic invariance of L ensures that there is a fact of the matter here. Then the following thesis is attractive:

Invariant Basis Thesis: There is an epistemically invariant language L such that every thought is implied by some sentence of L.

In fact, an apparently weaker thesis would suffice: every thought is implied by some epistemically invariant sentence. We need not require that all the sentences be part of the same language. But given the plausible thesis that for any class of languages, their union constitutes a language, the weaker principle entails the stronger.

I will not defend this thesis at length here. I think, however, that it (or something like it) is plausible. The central intuition behind it is that there is nothing about the *variability* of language that increases our ability to entertain maximal thoughts: variability simply allows for a certain sort of coarse-graining. And where there is coarse-graining, there are more specific fine-grained states, which can be expressed in a fine-grained language. If so, it is plausible that even if a specific thought cannot be expressed in an epistemically invariant sentence, it will be implied by a more fine-grained thought that can.

(As earlier, the thesis needs qualification to handle demonstrative thoughts. To handle such thoughts, the language can be supplemented by an arbitrary number of experiential demonstratives, different

demonstratives for each subject.)

In any case, if the thesis is true, we can use it to straightforwardly construct epistemic space. The natural approach is to identify scenarios with equivalence classes of maximal sentences (epistemically possible sentences of L that are not implied by any sentences that they do not imply), and to say that a scenario verifies a thought when one of the associated maximal sentences implies the thought. As for the actualization relation, formal definitions are subject to the same complications as before, but we given that we have a reasonably firm intuitive grasp on the notion, we can simply assume it.

Is Plenitude satisfied? If every thought T is implied by a maximal thought M , then M will be implied by a sentence of L , which will be maximal (given that every sentence expresses a thought), so there will be a scenario verifying T . Even if there is a problem with maximal thoughts, one could construct scenarios from maximal complexes of sentences, assuming we have an epistemic possibility operator over classes of sentences (yielding notions of implication and maximality over complexes), and that any sentence is implied by some maximal complex. Any thought will be implied by some sentence of L , which will be implied by a maximal complex, so the thought will be verified by a scenario.

As for the other principles: Compositionality will be satisfied straightforwardly. Parsimony will be satisfied if every sentence of L can express a thought; if not, it may be satisfied under other minimal assumptions. Actuality and Truth will be unproblematic, given an appropriate actualization relation. Translatability will be guaranteed, if we take our initial notion of translatability to hold that two thoughts are translatable when they are expressible by the same sentence of L . So if the thesis is true, minor assumptions yield a space of scenarios and a verification relation satisfying the central principles of epistemic space.

It is worth noting that this general discussion of epistemic translatability gives us a general constraint on epistemic space that may be worth noting. The discussion here gives us one way of independently grounding a notion of epistemic translatability among at least a subclass of thoughts: two thoughts are epistemically translatable when they are expressible by the same sentence of an epistemically invariant language.

6 Applications

We have seen that there are various ways in which we might construct a space of scenarios satisfying the principles of epistemic space. As with many such constructions, certain assumptions and/or idealizations are required for the constructions to work and I have not demonstrated the truth of the assumptions and the reasonableness of the idealizations. But the assumptions and idealizations do not look obviously problematic. And as with other constructions of abstract objects, the notion of a scenario should not be seen as hostage to a construction. Rather, the constructions provide a *prima facie* reason to think that the notion of epistemic space is coherent, and give a sense of how it should behave.

If the notion of epistemic space is coherent, it has a large payoff. One obvious application is in the

analysis of knowledge and of epistemic possibility. But there are also applications in the analysis of the content of thought and language. I will briefly outline three such application in the following: to the analysis of Fregean sense, narrow content, and indicative conditionals. Each of these topics deserves a much extended treatment (see "On Sense and Intension", "The Components of Content", and "The Tyranny of the Subjunctive" for this treatment). But even a brief examination shows that epistemic space has much to offer here.

6.1 Fregean sense

Frege held that every expression was associated with a referent, and also with a sense. The sense of an expression mirrors the cognitive value of the expression: for example, since "Hesperus is Phosphorus" is cognitively significant, "Hesperus" and "Phosphorus" have different senses, although they have the same referent. Frege also held that sense determines reference. In a little more detail, sense was supposed to obey at least the following four principles:

- (1) Every expression with a referent has a sense.
- (2) Sense reflects cognitive significance: singular terms 'a' and 'b' have different senses if 'a=b' is cognitively significant; sentences 'S' and 'T' have different senses iff 'S iff T' is cognitively significant.
- (3) The sense of a complex expression is determined by the senses of its parts.
- (4) Sense determines reference.

In contrast to Frege, many contemporary philosophers hold that there is no such thing as sense: nothing satisfies theses (1)-(4).

Epistemic space yields a notion that is closely akin to sense. We have seen that every statement can be associated with an epistemic intension, a function from scenarios to truth-values. Further, every expression token within a statement, can be associated with an epistemic intension, a function from scenarios to extensions. To unify these, let us say that the extension of a statement is its truth-value. Let us say that a statement is epistemically contingent if its negation is epistemically possible. Then epistemic intensions has the following properties:

- (1) Every expression token with an extension has an epistemic intension.
- (2) Epistemic intension reflects (deep) epistemic contingency: singular terms 'a' and 'b' have different epistemic intensions iff 'a=b' is epistemically contingent; statements 'S' and 'T' have different epistemic intensions iff 'S iff T' is epistemically contingent.
- (3) The epistemic intension of a complex expression is (usually) determined by the epistemic intension of

its parts.

(4) An expression's epistemic intension determines its extension, in conjunction with the actual world.

Principle (1) arises from the general framework. Principle (2) follows from Plenitude and Compositionality. Principle (3) follows from Compositionality. Principle (4) follows from Actuality and Truth.

These four principles are closely related to Frege's, with epistemic intensions playing the role of sense. In (1), senses are associated with expression tokens rather than types, but this seems also to be Frege's view at least of natural languages. In (3), we may need to put minor restrictions on compositionality of epistemic intensions (for reasons below), but it is close to correct. In (4), we must understand determination as determination in conjunction with the world: roughly, that sense provides a *condition* on extension, which entities in the world may satisfy. This may or may not have been Frege's view of determination, but in any case his view is often understood this way.

The key principle (2) will correspond to Frege's principle if cognitive significance is understood as epistemic contingency. Epistemic contingency behaves very much like cognitive significance: for example, my statement "Hesperus is Phosphorus" is epistemically contingent while "Hesperus is Hesperus" is not. More generally, any a posteriori statement will be both cognitively significant and epistemically contingent, while any trivial statement will be neither. There is a difference, however, due to the idealization in the notion of epistemic possibility. Frege held that many a priori statements, such as statements of mathematics, are cognitively significant. As I have defined (ideal deep) epistemic possibility, they are not epistemically contingent but epistemically necessary. It follows that this notion of epistemic contingency departs from Frege's notion of cognitive significance, as it is idealized and less fine-grained. It is nevertheless closely related, and can do much of the work done by cognitive significance. (It is also possible that we can define more fine-grained notions of epistemic possibility and epistemic intension, as we will see.)

In recent years, it has been argued that no broadly Fregean account of the content of linguistic expressions can work. There have been at least three prominent arguments here: the argument from indexicality (Perry), the modal argument (Kripke), and the epistemic argument (Kripke). It can easily be seen that these objections do not apply to the present account.

The argument from indexicality holds that there is no coherent way for a Fregean view to handle the sense indexicals such as "I" and "now". Frege held that these vary between speakers and/or occasions, but this yields a mismatch between sense and cognitive significance. On my approach, these are handled straightforwardly: each token of such an indexical has the same (indexical) epistemic intension, mapping a scenario to the individual or time at the center. This way, sense mirrors cognitive significance. This view departs from Frege's own view: because the sense of an indexical sentence is itself indexical, it will not have an absolute truth-value (as Frege held), but rather a truth-value relative to a subject and time. But the departure ensures consistency, and is still compatible with the spirit of a broadly Fregean view.

The modal argument suggests that a name is not semantically equivalent to any description because of differences in modal contexts. For example, it might have been that Hesperus was not the evening star (e.g. it had been knocked off course by a comet), so 'Hesperus' is not equivalent to 'the evening star'. And it seems that these cannot have the same sense, since they have different referents in possible worlds. On my view: once we distinguish epistemic from subjunctive notions, all this is compatible with 'Hesperus' and 'the evening star' having the same sense. 'Hesperus is not the evening star' is subjunctively possible but (roughly) epistemically impossible. Correspondingly, 'Hesperus' and 'the evening star' have the same epistemic intension (sense), but different subjunctive intensions. The modal context 'it might have been that ...' is governed by the subjunctive intensions of the expressions, as is the referent of these expressions in alternative possible worlds (considered as ways the world might have been). But all this is compatible with sameness of sense, where sense reflects cognitive significance and determines actual extension.

The epistemic argument suggests that many names are not equivalent to descriptions, because of epistemic differences. For example, 'Gödel' is not equivalent to 'the man who proved the incompleteness of arithmetic': it might turn out that Gödel did not prove incompleteness (perhaps he stole the proof), so it is not a priori that Gödel proved incompleteness. So the name and the description are inequivalent and have different senses. In response: nothing here can show that the name lacks an epistemic intension. In fact, Kripke's argument proceeds by considering an epistemically possible scenario (one in which the man called 'Gödel' steals the proof from a man called 'Schmidt'), and evaluating it as an instance of the epistemic possibility that Gödel did not prove incompleteness. All this is entirely compatible with the epistemic model. At most, it shows that the epistemic intension of a name is not equivalent to that of any associated description; it does not show that names do not have epistemic intensions.

So it seems that the framework of epistemic space provides the basis for a workable Fregean approach to the content of linguistic expressions. These arguments simply suggest that senses must be indexicals, that sense and extension do not exhaust semantic value, and that senses are intensions rather than descriptions. The result differs in some respects from Frege's own view (particularly with regard to the indexicality of senses, and the cognitive significance of a priori statements), but it clearly has much in common.

It should be noted that because epistemic intensions are associated in the first instance with expression tokens rather than types, this does not constitute an account of senses as linguistic meanings: the sort common to all tokens of an expression type. It may be that in some cases (but not all), it could be extended into such an account. But in any case, Frege himself allows that senses could vary between tokens of a type in natural language; and there is much work that a token-relative notion of linguistic content can do.

6.2 Narrow Content

Let us say that the content of a thought is *wide* if it depends on the environment: that is, if subjects who are internal (physical and phenomenal) duplicates have corresponding thoughts with different content.

The content of a thought is *narrow* if it does not depend on the environment, and is the same between duplicates. In contemporary philosophy of mind, it is commonly accepted that many thoughts have only wide content. In response, some have argued that thoughts also have narrow content, and that even when a thought has wide content, it also has narrow content underlying. But no account of narrow content has received widespread support.

I think that the epistemic content of a thought, as I have defined it, is a sort of narrow content. The corresponding thoughts of two internal duplicates will have the same epistemic content. This can be seen initially though examining the cases that are commonly used as arguments for wide content.

(1) *Twin Earth* (Putnam 1975). Oscar lives on Earth, surrounded by water (i.e. H₂O). Twin Oscar lives on Twin Earth, which is much like Earth except that H₂O is replaced by superficially identical XYZ, which is not water but twater. Oscar and Twin Oscar are (near) duplicates and are chemically ignorant. Both utter "water is expensive", expressing a belief. Oscar believes that water is expensive, but Twin Oscar does not: he believes that twater is expensive. So their beliefs have different content.

Let B1 and B2 be the respective beliefs of Oscar and Twin Oscar. What is their epistemic content? We can see this by examining four relevant scenarios:

S1: Watery stuff in the environment is H₂O, H₂O is expensive. S2: Watery stuff in the environment is H₂O, H₂O is not expensive. S3: Watery stuff in the environment is XYZ, XYZ is expensive. S4: Watery stuff in the environment is XYZ, XYZ is not expensive.

Oscar's belief B1 is clearly verified by S1, and falsified by S2. It is also verified by S3: the thought that S3 is actual implies B1. That is, for Oscar, the hypothesis that S3 is actual and water is not expensive is epistemically impossible: if he accepts that S3 is actual, he will rationally accept that water is expensive. For similar reasons, B1 is falsified by S4.

For symmetrical reasons, Twin Oscar's belief B2 is verified by S3 and falsified by S4, and verified by S1 and falsified by S2. That is, both B1 and B2 are verified by S1 and S3, and falsified by S2 and S4. So at least as far as these four scenarios are concerned, B1 and B2 have the same epistemic intension. But these four scenarios are the crucial scenarios here: if B1 and B2 have the same epistemic intensions across these scenarios, they plausibly have the same epistemic intension overall. So B1 and B2 plausibly have the same epistemic content.

Note that B1 and B2 plausibly have different *subjunctive* content. B1 is plausibly *satisfied* only by worlds in which H₂O is expensive and not by worlds in which XYZ is expensive, while for B2, the situation is reversed. So B1 and B2 have different subjunctive intensions. The subjunctive content here can be seen as the wide content of B1 and B2, while the epistemic content can be seen as the narrow content.

(2) *Arthritis* (Burge 1979). Bert and Twin Bert both have a term "arthritis", which they use for a disease whose nature they do not fully understand. Bert lives on Earth, where experts in his community use

"arthritis" for disease D1 (arthritis), Twin Bert lives on Twin Earth, where experts in his community use "arthritis" for disease D2 (twarthritis). Bert and Twin Bert are duplicates: both utter "I have arthritis in my thigh". Bert believes that he has arthritis in his thigh. Twin Bert believes that he has twarthritis in his thigh. So it seems that their beliefs have different content.

Let B1 and B2 be the respective beliefs of Oscar and Twin Oscar. What is their epistemic content? We can see this by examining four relevant scenarios:

S1: Experts use "arthritis" for D1, D1 is in my (center subject's) thigh. S2: Experts use "arthritis" for D1, D1 is not in my thigh. S3: Experts use "arthritis" for D2, D2 is in my thigh. S4: Experts use "arthritis" for D2, D2 is not in my thigh.

(Scenario S1 may be impossible due to the nature of D1, but set this inessential worry aside.) Oscar's belief B1 is clearly verified by S1 (the actualized scenario) and falsified by S2. It is also verified by S3: the thought that S3 is actual implies B1. That is, for Oscar the hypothesis "S3 is actual and arthritis is not in my thigh" is epistemically impossible: if Oscar accepts that S3 is actual, he should rationally accept "I have arthritis in my thigh". For similar reasons, B1 is falsified by S4.

For symmetrical reasons, Twin Bert's belief B2 is verified by S3 (his actualized scenario) and falsified by S4, and verified by S1 and falsified by S2. That is, both B1 and B2 are verified by S1 and S3, and falsified by S2 and S4. So at least as far as these four scenarios are concerned, B1 and B2 have the same epistemic intension. But these four scenarios are the crucial scenarios here: if B1 and B2 have the same epistemic intensions across these scenarios, they plausibly have the same epistemic intension overall. So B1 and B2 plausibly have the same epistemic content.

Again, B1 and B2 may have different subjunctive intensions, so their subjunctive content may constitute wide content. But their epistemic content constitutes narrow content.

All this enables us to reconcile the facts that (a) the truth of a belief ascription such as "Oscar believes that water is expensive" or "Bert believes that he has arthritis in his thigh" depends on the subject's environment, and that (b) the two corresponding beliefs have the same narrow (epistemic) content. To reconcile these, we need only accept that the truth of belief ascription can depend on *both* the subjunctive content and the epistemic content of a subject's beliefs. Given this dual dependence, the wideness of subjunctive content explains the wideness of belief ascriptions, in a manner that is entirely compatible with epistemic content being narrow. (See "The Components of Content" for more on how this dual dependence works.)

In addition to examining cases, we can also give a principled argument that epistemic content is narrow. All we need are the following two theses:

Duplication Thesis 1: Given any two duplicates A1 and A2 and corresponding thoughts T1 and T2, then T1 is epistemically possible iff T2 is epistemically possible.

Duplication Thesis 2: Given any two duplicates A1 and A2 and corresponding concepts C1 and C2, then if C1 is an epistemically translatable basis concept, C2 translates C1.

The first thesis is extremely plausible, and nothing in arguments for externalism gives reason to deny it. In effect, the claim is that the apriority of beliefs is internally determined. This is what we would expect: whether a belief is a priori depends only on whether it can be nonexperientially justified, and whether a belief can be nonexperientially justified is independent of the environment.

The second thesis is also plausible, for the epistemic translation relation invoked earlier. The sorts of concepts over which translation is invoked are concepts to which externalist arguments do not apply: extension-dependent concept types such as names and natural kind concepts are excluded, and deferential uses of concepts are excluded. For the indexical and purely descriptive concepts that remain, the translation mapping between duplicate individuals will be relatively uncontroversial. And if these concepts form a basis class, this mapping can be extended into a mapping of arbitrary scenarios.

Given these two theses, then take any two duplicates with corresponding beliefs B1 and B2, and consider an arbitrary scenario S. Given that the translatable concepts are a basis class, S will be associated with a translatable maximal thought M1 in the first subject. If S verifies B1, then M will imply B1: that is, the thought "M1&~B1" will be epistemically impossible. The thought corresponding to "M1&~B1" will be "M2&~B2". By the first thesis, M2&~B2 will be epistemically impossible. By the second thesis, M2 will translate M1, so it will also be associated with S. So we can see that S verifies B1 if and only if S verifies B2. So B1 and B2 have the same epistemic content. This holds for arbitrary beliefs, so epistemic content is narrow content.

(As before, there is a mild exception for demonstrative beliefs, whose content is *sui generis* between individuals. When I attend to a spot in a symmetrical visual field and believe "This is red", and when a duplicate of me does the same, our thoughts will not have exactly the same epistemic content, since there is no canonical translation between experiential demonstratives. Instead, we can say that the thoughts are *epistemically isomorphic*, and have isomorphic epistemic content. Given the general thesis that corresponding demonstrative concepts are epistemically isomorphic concepts, as is plausible, we can still maintain the thesis that duplicates always have isomorphic narrow contents.)

6.3 Indicative and subjunctive conditionals

There is a striking, although infrequently noted asymmetry between indicative and subjunctive conditionals. Consider the following pairs:

(1a) If Prince Albert Victor murdered those women, he is Jack the Ripper.

(1b) If Prince Albert Victor had murdered those women, he would have been Jack the Ripper.

and

(2a) If XYZ is (and has been) the clear liquid in the oceans and lakes, water is XYZ.

(2b) If XYZ had been the clear liquid in the oceans and lakes, water would have been XYZ.

In these pairs, the first member seems intuitively correct. (1a) might reasonably be accepted by a detective investigating the murders. (2a) might reasonably be accepted by someone who is ignorant of the chemical composition of water, or even by someone who knows that water is H₂O but is not certain of this. These conditionals pass the Ramsey test: if the antecedent is accepted as a hypothetical premise, the consequent should be accepted as a hypothetical conclusion. Let us say that in these conditions, the indicative conditional is *correct*. (I will be neutral on whether correctness corresponds to truth, assertibility, or something else.) Clearly, (1a) and (2b) are correct indicative conditionals.

In these pairs, the second member seems intuitively incorrect, at least if Kripke's intuitions are accepted. Given that Prince Albert Victor is not actually Jack the Ripper (as seems likely), then even if he had committed the murders, he would not have been Jack the Ripper. Rather, he would have committed the murders *instead* of Jack the Ripper. Similarly, given that water is actually H₂O, then even if XYZ had been the clear liquid in the oceans and lakes, XYZ would not have been water. Rather, XYZ would have been in the oceans and lakes *instead* of water.

(I should note that my own intuitions about these subjunctive conditionals are not quite as clear as Kripke's. But for the sake of discussion I will accept the Kripkean intuitions.)

There is one especially striking feature of the asymmetry: it seems that the correct indicative conditionals have a "metaphysically impossible" consequent. How can this be accommodated in a semantic theory? More generally, how can the asymmetry be explained?

Many philosophers hold that one can give a possible worlds account of subjunctive conditionals, along something like the following lines:

"If it had been that S, it would have been that T" is true $\leftrightarrow$ the closest S-world is a T-world.

Here, closeness is a similarity metric on the space of possible worlds, and the closest S-world is the world such that of all worlds in which S in the case, it is the most similar to the actual world. Note that the closeness metric may vary between contexts, and that one can adjust the definition in various ways, e.g. to allow for multiple closest worlds.

Many philosophers hold that by contrast, one cannot give a possible-worlds account of indicative conditionals, holding that they are deeply epistemic rather than modal. The considerations above are not usually invoked here, but they might be thought to strengthen the case: if correct indicative conditionals can have metaphysically possible antecedents and impossible consequences, then possible worlds are an

inappropriate tool.

However, the framework of epistemic space allow us to give an analysis of indicative conditionals that is parallel to the analysis of subjunctive conditionals, except that we invoke the space of epistemic rather than subjunctive possibility. In particular, one can hold the following

"If S is the case, then T is the case" is correct $\leftrightarrow$ the epistemically closest S-scenario is a T-scenario.

I will not attempt an exact analysis of the notion of epistemic closeness here. But it should be understood as something like the following: one scenario is epistemically closer than another (for a subject) to the extent that it is more compatible with the subject's knowledge (or beliefs), or to the extent that it is compatible with a smaller revision of the subject's knowledge-constituting beliefs. Note that epistemic closeness is subject-relative, and depends on what a subject knows (or believes). Whether knowledge or belief should be invoked here is a tricky issue that I will not go into here.

Then in the cases above: there is a wide variety of scenarios in that verify "Prince Albert Victor committed the murders", and all (or almost all) verify "Prince Albert Victor is Jack the Ripper". Similarly, there is a wide variety of scenarios in which XYZ is and has been the liquid in the oceans and lakes. Those most compatible with a subject's beliefs are plausibly those in which things are superficially identical except for the substitution for XYZ for H₂O, and these all verify "water is XYZ". So in both cases, the epistemically closest scenarios verifying the antecedent also verify the consequent.

Note that all this is very much compatible with the Ramsey test for the assertibility of indicative conditionals. When we hypothetically accept S, we hypothetically exclude all scenarios verifying $\sim$ S, while making minimal revisions to our beliefs. The scenarios we will then be led to hypothetically endorse are those that verify $\sim$ S and that are compatible with a minimal revision of our beliefs. That is, we are led to hypothetically endorse the epistemically closest S-scenarios. We will conclude that T precisely to the extent to which the endorsed S-scenarios are also T-scenarios. So the Ramsey test will be satisfied precisely when the conditions above are met.

(Note that the Ramsey test corresponds to the belief-relative notion of epistemic closeness. Does this mean that our correctness conditions should also appeal to the belief-relative notion? Not necessarily: or only to the extent that correctness and assertibility converge. It is arguable that when an assertible indicative conditional is grounded in false beliefs, the conditional is intuitively incorrect. If so, then it is plausible that correctness conditions should not depend on a subject's false beliefs, and perhaps should depend only on a subject's knowledge.)

The subject-relativity in the notion of epistemic closeness mirrors the often-noted subject-relativity in the correctness conditions of indicative conditionals. To use a version of the "Sly Pete" example: say that Pete holds one card, that I know that Pete has a 3 or a 5, and that you know that Pete has a 5 or a 7 (of course he actually has a 5). Then I can correctly say: if Pete does not have a 5, he has a 3. And you can correctly say: if Pete does not have a 5, he has a 7. It is plausible that I could not correctly assert the

second, and you could not correctly assert the first. The difference results from the difference in our epistemic closeness measure, resulting from the difference in our knowledge. For me, the closest scenarios in which Pete does not have a 5 are scenarios in which he has a 31. For you, the closest scenarios in which Pete does not have a 5 are those in which he has a 7.

Note that the analyses of subjunctive and indicative conditionals are almost exactly parallel, except for the differences between (i) S-worlds vs. S-scenarios, and (ii) counterfactual closeness vs. epistemic closeness. It is natural to hope that many of the tools that have been used to analyze subjunctive conditionals can also be adapted to analyze indicative conditionals.

7 Hyperintensionality

The notion of (deep) epistemic possibility that we have been dealing with is an idealized one: if S is knowable a priori, then it is not epistemically possible that $\sim S$, even when S is far from obvious, and even when no-one in the world knows that S. This has the result that any two sentences that are a priori equivalent will have the same epistemic intension, and the same epistemic content. For example, all a priori sentences, including complex sentences of logic and mathematics, have the same epistemic intension. And all a priori thoughts, including complex logical and mathematical thoughts, have the same epistemic intension.

It is natural to wonder if there is a less idealized notion of epistemic possibility that will a more fine-grained notion of epistemic content. This is a complex issue, but I think there can be.

To develop this notion, we must start with a *nonideal* notion of deep epistemic possibility. Instead of saying that it is epistemically possible that P when $\sim P$ cannot possibly be ruled out a priori, we might say that it is epistemically possible that P when $\sim P$ cannot be ruled out *through reasoning of a certain sort*. Here, there are various possibilities.

We might hold that it is epistemically possible that P when:

(i) it is not obvious a priori that $\sim P$; (ii) P cannot be ruled out by such-and-such amount of a priori reasoning; (iii) P cannot be ruled out through logical reasoning alone; (iv) P cannot be ruled out through non-moral a priori reasoning; (v) it is not obvious that $\sim P$ (vi) it is not trivial that $\sim P$ (vii) it is not certain that $\sim P$ (viii) it is not known that $\sim P$

and so on. Here, the definitions work have different properties. The first four definitions all appeal to varieties of a priori reasoning. These guarantee that if it is epistemically possible that P in the previous sense, it will be epistemically possible that P in the new sense. The last four definitions do not appeal only to a priori reasoning; so it is possible that some P will be epistemically possible in the original sense but not in the new sense (e.g. obvious empirical falsehoods, such as "I do not exist" or "I do not have hands"). Because we are mostly concerned with more fine-grained notions of epistemic possibility here, I will concentrate on definitions in the first class.

Given a notion of non-ideal epistemic possibility, we can attempt to set up a corresponding *non-ideal* epistemic space, made up of *non-ideal* scenarios. The principles governing this space will be much as before. The key principle, once again, will be Plenitude: there is a scenario verifying S iff S is epistemically possible. Because many more thoughts and statements will be epistemically possible for non-ideal notions of epistemic possibility, it follows that there will be many more corresponding non-ideal scenarios.

It seems reasonable that the principles of Actuality, Truth, Parsimony, and Translatability should all hold on this model. There is a question about whether Compositionality should be endorsed. If a compositional principle is itself nonobvious, it may be that each T_i can be epistemically possible without $\text{comp}(T_i)$ being epistemically possible. If compositionality holds, it is likely that if a scenario verifies some statements, it will verify all logical consequences of those statements, which may be undesirable. If so, we may wish to do without Compositionality, or restrict it in some fashion to obvious compositions and the like.

The processing of *constructing* scenarios will be more complex where non-ideal epistemic possibility is concerned. It is clear that centered worlds will be inappropriate here. On natural models of verification, a priori falsehoods will be verified by no centered world; and on any reasonable model of verification, there will not be enough centered worlds to satisfy Plenitude. As for the direct epistemic construction, we will probably need to avoid appeal to maximal statements and thoughts. Maximal statements and thoughts are likely to be indiscriminately epistemically possible, because of their complexity, and they may have no interesting implication relations to ordinary thoughts.

Instead, it may be best to appeal to *classes* of statements of thoughts: perhaps classes such that no statement or combination of statements is epistemically impossible. Then we can say that a class implies a statement T if some conjunction of $\sim T$ with a set of statements in the class is epistemically impossible, or perhaps if there is some reasoning process of the relevant sort that takes us from a statement or group of statements in the class to T. We can say that one class implies another class if it implies every statement in that class. We can say that a class is maximal if it is implied by no class that it does not imply. A maximal class will verify a thought when it implies that thought. There will be difficulties in setting up equivalence relations on maximal classes, due to failures of transitivity in implication, but this problem might be dealt with in a variety of ways. It seems that this sort of approach is at least promising.

If we can set up a non-ideal epistemic space corresponding to a non-ideal notion of epistemic possibility, we will then have a corresponding notion of non-ideal epistemic content. We can say that the non-ideal epistemic content of a thought is the thought's intension over non-ideal scenarios, according to whether those scenarios verify the thought. Then for any two thoughts T1 and T2 such that it is epistemically possible that T1 holds without T2 and vice versa, T1 and T2 will have different non-ideal epistemic content. The same goes for the non-ideal epistemic content of statements.

When this way of thinking is applied to different notions of epistemic possibility, it will yield various different applications. For example, if we are concerned with Frege's notion of cognitive significance, we

can say that T is epistemically possible when $\sim T$ is cognitively significant (perhaps this will be whenever $\sim T$ is nontrivial), and we can set up a corresponding non-ideal epistemic space. This will yield a variety of non-ideal epistemic content that behaves very much like Fregean sense.

Another application: For the ordinary notion of epistemic possibility (strict epistemic possibility) with which we started this paper, it is plausible that P is strictly epistemically possible when one could not *easily* come to know that $\sim P$ given what one already knows. The corresponding notion of deep epistemic possibility is something like the following: it is deeply epistemically possible that P when one cannot easily know a priori that $\sim P$. From this notion, we will be able to set up a corresponding non-ideal epistemic space. For this space, it is plausible that P is strictly epistemically possible iff there is a P -scenario that is not excluded by any item of one's knowledge. So this space is perhaps the closest to delivering the intuitive picture of strict epistemic possibility as discussed at the start of the paper.

Another application: say that we are concerned with hypotheses about the relationship between the nonmoral and the moral. We may think that the connection is ultimately a priori, or we may think that moral beliefs are ultimately not truth-evaluable, but as long as the connection and the non-truth-evaluability is not obvious, there will be an interesting hypothesis space to investigate. To do this, we can invoke notion (iv) above: it is epistemically possible that P when P cannot be ruled out through nonmoral a priori reasoning. This will plausible yield a space of "moral scenarios" which is much like the space of ideal scenarios, except that it may have an additional dimension of variation in the way that it associates moral claims with nonmoral claims. Such a space may be very useful for analyzing the moral domain without presupposing moral views. (This will be closely related to Gibbard's notion of a factual-normative world.)

It may be that there is no canonical notion of nonideal epistemic possibility. If so, there will be no canonical notion of nonideal epistemic content. Instead, we might have a spectrum of notions of deep epistemic possibility, from the ideal to the nonideal, perhaps ending at the notion on which anything is epistemically possible, and on which there are no interesting relations among epistemic content. There will be a corresponding spectrum of epistemic spaces. Every statement might then be associated with a spectrum of epistemic intensions, each of which is an intensions across scenarios within a given epistemic space. For different purposes, different intensions from within this spectrum may be relevant. Between these intensions and these epistemic spaces, there will be enough material to do significant explanatory work in many different epistemic domains.

The Foundations of Two-Dimensional Semantics

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

*[[A first draft. All comments are extremely welcome.]]

1 Reason, meaning, and possibility

Why is two-dimensional semantics important? One can think of it as the fourth act in a triangular drama involving three of the central concepts of philosophy: reason, meaning, and possibility. In the prologue, Kant linked reason and possibility. In the first act, Frege proposed a link between reason and meaning. But it was not obvious how this link was to be understood. In the second act, Carnap linked meaning to possibility. In conjunction with the Kantian link between reason and possibility, this built a Fregean bridge from reason to meaning. The result was a golden triangle of constitutive connections between meaning, reason, and possibility. In the third act, Kripke severed the link between reason and possibility. The link between meaning and possibility was left intact, but the bridge between meaning and reason was broken.

Two-dimensional semantics promises to restore the golden triangle. While acknowledging the aspects of meaning and possibility that derive from the third act, it promises to explicate further aspects of meaning and possibility that are more closely tied to reason. In particular it promises to look at possibility and the evaluation of possibilities in a different way, and erect a notion of meaning on that basis. In this way, the constitutive connections between meaning, reason, and possibility might be restored.

To date, the restoration has been incomplete. Many different frameworks for two-dimensional semantics have been proposed, and many of them restore the triangle at best partially. It is controversial whether two-dimensional semantics can be understood in such a way that the triangle is fully restored. To see whether a restoration is possible, we need to investigate the foundations of two-dimensional semantics, and explore the many different ways in which the framework can be understood. I think that when the framework is understood in the right way, it can reinstantiate the links between meaning, reason, and possibility.

1.1 Frege, Carnap, and Kripke

We can begin with some background, in the form of a simplified rational reconstruction of history.

First, Frege. Frege held that an expression in a language typically has a *referent* - or what I will here call an *extension*. The extension of a singular term is an individual: for example, the extension of the name 'Hesperus' is the planet Venus, and the extension of the description 'the teacher of Aristotle' is Plato. The extension of a general term is a class. And the extension of a sentence is its truth-value.

Frege noted that the extension of an expression does not in general determine its cognitive significance: the role it plays in reasoning and in knowledge. For example, 'Hesperus' (the name used for the evening star) and 'Phosphorus' (the name used for the morning star) have the same referent but have different cognitive significance, as witnessed by the fact that 'Hesperus is Hesperus' is cognitively trivial, while 'Hesperus is Phosphorus' is nontrivial. The same goes for many other pairs of expressions: perhaps 'renate' (creature with a heart) and 'cordate' (creature with a kidney), or 'water' and 'H₂O', or 'I' (as used by me) and 'David Chalmers'. In each pair, the members are co-extensive (they have the same extension), but they are cognitively and rationally distinct.

Frege held that meaning is tied constitutively to cognitive significance, so that if two expressions have different cognitive significance, they have different meaning. It follows that there must be more to meaning than extension. Frege postulated a second aspect to meaning: *sense*. When two expressions are cognitively distinct, they have different senses. For example, the nontriviality of 'Hesperus is Phosphorus' entails that although 'Hesperus' and 'Phosphorus' have the same extension, they have a different sense. We can put the general idea as follows:

Fregean Thesis: Two expressions 'A' and 'B' have the same sense iff 'A == B' is cognitively insignificant.

Here, 'A == B' is a claim that is true if and only if 'A' and 'B' have the same extension. Where 'A' and 'B' are singular terms, this will be the identity 'A=B'; where 'A' and 'B' are sentences, this will be the material biconditional 'A iff B'; and so on. As for cognitive significance, we can say at a first approximation that a claim is cognitively insignificant when it can be known trivially by a rational being. As such, we can see this characterization of sense as providing a first bridge between meaning and reason.

The idea that expressions have senses is attractive, but senses are nevertheless elusive. What exactly is a sense? What is cognitive significance? How does one analyze meanings beyond extensions? In the middle part of the 20th century, a number of philosophers, most notably Carnap, had an insight. We can use the notions of *possibility* and *necessity* to help understand meaning, and in particular to help understand sense.

There are many possible ways the world might be; and we can use language to describe these possibilities. An expression can be applied to the actual state of the world, yielding an actual extension, or it can be applied to alternative possible states of the world, yielding alternative possible extensions. Take expressions such as 'renate' and 'cordate'. In the world as it actually is, all renates are cordates, so these terms have the same extension. But it is not *necessary* that all renates are cordates: if the world had been

different, some renates might have failed to be cordates. Applied to such an alternative possibility, the two terms have a different extension. We can say: 'renate' and 'cordate' are co-extensive, but they are not *necessarily* co-extensive. Carnap suggested that we say two expressions have the same *intension* if and only if they are necessarily co-extensive.[*] So 'renate' and 'cordate' have the same extension, but different intension. We can put the general claim as follows:

Carnapian Thesis: 'A' and 'B' have the same intension iff 'A == B' is necessary.

*[[See Carnap 1947. The idea is also present in Lewis 1944.]]

What exactly is an intension? Carnap's characterization suggests a natural definition: an intension is a function from possibilities to extensions.[*] The space of possibilities here correspond to different possible states of the world. Relative to any possibility, an expression has an extension: for example, a sentence (e.g. 'all renates are cordates') can be true or false relative to a possibility, and a singular term (e.g. 'the teacher of Aristotle') picks out an extension relative to a possibility. An expression's intension is the function that maps a possibility to the expression's extension relative to that possibility. When two expressions are necessarily co-extensive, they will pick out the same extension relative to all possibilities, so they will have the same intension. When two expressions are not necessarily co-extensive, they will have not pick out the same extension relative to all possibilities, so they will have different intensions. So intensions behave just as Carnap suggests they should.

*[[Although this definition of an intension is often attributed to Carnap, I do not see it explicitly in Carnap 1947. [Citations welcome!] Although he characterizes necessity ("L-truth") initially in terms of state-descriptions, which are akin to possible worlds, state-descriptions very soon drop out of the discussion, so that intensions are treated in effect as something of a primitive semantic value. Lewis's suggestion that an intension "comprises whatever must be true of any possible world in order that the proposition should apply to it or be true of it" suggests at least that intensions determine functions in the above sense, however.]]

Seen this way, the notion of an intension provides a bridge between meaning and possibility. Just as a sense can be seen as a sort of meaning that is constitutively tied to reason, an intension can be seen as a sort of meaning that is constitutively tied to possibility. Furthermore, intensions seem to behave very much as senses are supposed to behave. Just as two expressions can have the same extension but different senses, two expressions can have the same extension but different intensions. And just as sense was supposed to determine extension, intension seems to determine extension, at least relative to a world.

One can make a direct connection by adding an additional claim connecting possibility and reason. It has often been held that a proposition is necessary if and only if it is a priori (knowable by reason alone) or trivial (yields no real knowledge of the world). The notions of apriority and triviality are essentially rational notions, defined in terms of how one comes to know a claim. Carnap himself held a version of the thesis involving triviality, but it is more useful for our purposes to focus on the version involving apriority. In this form, the relevant thesis goes back at least to Kant, so we can call it:

Kantian Thesis: A proposition P is necessary iff P is a priori.

If we combine the Carnapian Thesis with the Kantian Thesis, we obtain the following:

Neo-Fregean Thesis: 'A' and 'B' have the same intension iff 'A == B' is a priori.

If this claim is accepted, then one has recaptured something that is at least close to the Fregean thesis. For apriority is at least closely related to cognitive insignificance. When a proposition is cognitively insignificant, it is plausibly a priori. The reverse is not quite the case, as Frege understood cognitive significance: many logical and mathematical propositions are cognitively significant, even though they are a priori. But in any case, a priority and cognitive insignificance are at least closely related rational notions. Typical cognitively significant identities, such as 'Hesperus is Phosphorus', 'water is H₂O', and 'I am David Chalmers' are all a posteriori. If the Neo-Fregean Thesis is correct, it follows that 'Hesperus' and 'Phosphorus' have different intensions, as do 'water' and 'H₂O', and 'I' and 'David Chalmers'. So intensions behave quite like Fregean senses.

In effect, the notion of possibility serves as a bridge in explicating the tie between meaning and reason. One constructs a notion of meaning using the notion of possibility, combines this with the claim that possibility is constitutively tied to reason, and ends with a link between all three. The central connection between meaning, reason, and possibility is captured within the core thesis: intension is a notion of meaning, defined in terms of possibility, that is constitutively connected to reason.

This golden triangle was shattered by Kripke, who cut the connection between reason and possibility. Kripke argued that the Kantian Thesis is false: there are many propositions that are necessarily true but not knowable a priori. For example, Kripke argued that given that Hesperus is actually Phosphorus, it could not have been that Hesperus was not Phosphorus: Hesperus is necessarily the planet Venus, and so is Phosphorus. So although 'Hesperus is Phosphorus' is not knowable a priori, it is nevertheless necessary. More generally, Kripke argued that names and natural kind terms are rigid designators, picking out the same extension in all possible worlds. It follows that any true identity involving such terms is necessary. For example, 'water is H₂O' is necessary, even though it is a posteriori. The same goes for claims involving indexicals: 'I am David Chalmers' (as used by me) is another a posteriori necessity.

If Kripke is right about the Kantian Thesis, then the Neo-Fregean Thesis is also false. Since 'Hesperus is Phosphorus' is necessary, 'Hesperus' and 'Phosphorus' have the same intension, picking out the planet Venus in all possibilities. But the equivalence between 'Hesperus' and 'Phosphorus' is nevertheless a posteriori and cognitively significant. So cognitively and rationally distinct pairs of expressions can have the same intension: witness 'Hesperus' and 'Phosphorus', 'water' and 'H₂O', 'I' and 'David Chalmers'. So the Neo-Fregean Thesis fails, and intensions no longer behave like Fregean senses.

In effect, Kripke leaves intact the Carnapian link between meaning and possibility, but in severing the Kantian link between reason and possibility, he also severs the Fregean link between meaning and reason. This is roughly the received view in contemporary philosophy: meaning and possibility are connected, but both are disconnected from reason.

1.2 Two-dimensional semantics

Although most contemporary philosophers accept Kripke's arguments against the Kantian thesis, many still would like to hold that Frege was right about *something*. There remains an intuition that 'Hesperus' and 'Phosphorus' (or 'water' and 'H₂O', or 'I' and 'David Chalmers') differ in at least some dimension of their meaning, corresponding to the difference in their cognitive and rational roles. One might try to do this by breaking the Carnapian connection between meaning and possibility. Two-dimensional semantics takes another strategy: in effect, it finds another way of looking at possibility that yields a more Fregean aspect of meaning.

The core idea of two-dimensional semantics is that there are two different ways in which the extension of an expression depends on possible states of the world. First, the actual extension of an expression depends on the character of the actual world in which an expression is uttered. Second, the counterfactual extension of an expression depends on the character of the counterfactual world in which the expression is evaluated. Corresponding to these two sorts of dependence, expressions correspondingly have two sorts of intensions, associating possible states of the world with extensions in different ways. On the two-dimensional framework, these two intensions can be seen as capturing two dimensions of meaning.

These two intensions correspond to two different ways of thinking of possibilities. In the first case, one thinks of a possibility as representing a way the actual world might turn out to be: or as it is sometimes put, one *considers a possibility as actual*. In the second case, one acknowledges that the actual world is fixed, and thinks of a possibility as a way the world might have been but is not: or as it is sometimes put, one *considers a possibility as counterfactual*. A version of this is commonly put by saying that one can think of a possible world as a *context of utterance*, or as a *circumstance of evaluation*. When one evaluates an expression relative to a possible world, one may get different results, depending on whether one considers the possible world as actual or as counterfactual.

The second way of thinking about possibilities is the more familiar in contemporary philosophy. Kripke's arguments rely on viewing possibilities in this way. Take a possibility in which the brightest object in the evening sky is a satellite around the earth, and in which Venus is visible and bright only in the morning. When we think of this possibility as a counterfactual way things might have been, we do not describe it as a possibility in which Hesperus is Mars, but as one in which Hesperus (and Phosphorus) is invisible in the evening. So relative to this possibility considered as counterfactual, 'Hesperus' picks out Venus. Correspondingly, the second-dimensional intensions of both 'Hesperus' and 'Phosphorus' both pick out Venus in this possibility, and in all possibilities in which Venus exists. It is this familiar sort of intension that yields the Kripkean gap between intension and cognitive significance.

The first way of thinking about possibilities is the less familiar in contemporary philosophy. If we take the possibility described above, and think of it as a way the world might actually be, we can say: if the world really is that way, then 'Hesperus' picks out a satellite. So relative to this possibility considered as actual, 'Hesperus' picks out not Venus but the satellite. Correspondingly, the first-dimensional intension of 'Hesperus' picks out the satellite in this possibility, while that of 'Phosphorus' picks out Venus. So

'Hesperus' and 'Phosphorus' have different first-dimensional intensions. This difference is tied to the fact that the actual-world reference of 'Hesperus' and 'Phosphorus' is fixed in quite different ways, although as things turn out, their referents coincide. Because of this, it seems that the first dimension may be better suited than the second for a link to reason and to cognitive significance.

The possibilities evaluated in the second dimension are usually thought of as possible worlds. The possibilities evaluated in the first dimension are a little different, as they reflect the nature of a world from the point of view of a speaker using an expression within a world. It is useful for many purposes to see these possibilities as *centered worlds*: worlds marked with a "center", which is an ordered pair of an individual and a time. We can think of the center of the world as representing the perspective of the speaker within the world.

I have been deliberately vague about just how the relevant intensions are to be defined, since as we will see, there are many different ways to define them. Because of this, giving detailed examples is tricky, because different frameworks treat cases differently. Nevertheless, it is useful to go through some examples, giving an intuitive analysis of the results that two-dimensional semantics might be expected to give *if* it is to yield something like a Fregean sense in the first dimension. We will later see how this can be cashed out in detail. For now, I will use "1-intension" as a generic name for a first-dimensional intension, and "2-intension" as a generic name for a second-dimensional intension.

First, 'Hesperus is Phosphorus'. In a centered world considered as actual, this is true roughly when the morning star visible from the center of that world is the same as the evening star. In a world considered as counterfactual, it is true when Venus is Venus. 'Hesperus' functions roughly to pick out the evening star in the actual world, so the 1-intension of 'Hesperus' picks out the evening star in a given centered world. Likewise, the 1-intension of 'Phosphorus' picks out the morning star in a centered world. Both of these terms behave rigidly in counterfactual evaluation, so their 2-intensions pick out their actual referents in all worlds. So the 2-intensions of both 'Hesperus' and 'Phosphorus' pick out Venus in all worlds.

Second, 'water is H₂O'. In a centered world considered as actual, this is true roughly when the clear, drinkable liquid around the center of that world has a certain pattern of chemical structure. In a world considered as counterfactual, it is true when H₂O is H₂O. The reference of 'water' is fixed roughly by picking out the substance with certain superficial properties and a certain connection to the speaker in the actual world, so its 1-intension picks out roughly the substance with those properties connected to the center of a given world. Similarly, the 1-intension of 'H₂O' picks out the substance with the right sort of chemical structure in a centered world. As in the first case, both expressions behave rigidly in counterfactual evaluation, so their 2-intensions pick out H₂O in all worlds.

Third, 'I am a philosopher'. In a centered world considered as actual, this sentence is true when the being at the center of the world is a philosopher. In a world considered as counterfactual, this sentence (or at least my utterance of it) is true if David Chalmers is a philosopher in that world. The actual-world reference of 'I' is fixed by picking out the subject who utters the token; so the 1-intension of 'I' picks out the subject at the center of a given world. 'I' behaves as a rigid designator in counterfactual evaluation, so its 2-intension picks out the actual referent (in this case, David Chalmers) in all possible worlds.

'Philosopher', by contrast, is a broadly descriptive term: both its 1-intension and its 2-intension function to pick out beings with certain characteristic attributes.

Certain patterns seem to emerge. The first two sentences are necessary (at least if Kripke is right), and both of them have a 2-intension that is true in all worlds. The third sentence is contingent, and its 2-intension is false in some worlds. So it seems that a sentence is necessary precisely when it has a necessary 2-intension. This corresponds directly to the Carnapian thesis: 2-intensions, in effect, are defined so that two expressions will have the same 2-intensions when they are necessarily equivalent.

On the other hand, all three of these sentences are a posteriori, and all of them have a 1-intension that is false in some centered worlds. At the same time, a priori sentences such as (perhaps) 'All bachelors are unmarried males' or 'Hesperus (if it exists) is visible in the evening' can plausibly be seen as having a 1-intension that is true in all centered worlds. So it is at least tempting to say that a sentence is a priori precisely when it has a necessary 1-intension. This corresponds to the neo-Fregean thesis: one might naturally suggest that two expressions have the same 1-intension precisely when they are a priori equivalent. To illustrate, one can note that the difference in the 1-intensions of 'Hesperus' and 'Phosphorus', or of 'water' and 'H₂O', seems to be closely tied to their a priori inequivalence. All this needs to be analyzed in more depth, but one might at least characterize the general sort of behavior suggested in the examples above, where 1-intensions go along at least roughly with differences in cognitive significance, as *quasi-Fregean*.

Along with the 1-intension and the 2-intension of a given expression, one can also define a *two-dimensional intension*. In many cases, just as an expression's extension depends on how the actual world turns out, an expression's 2-intension depends on how the actual world turns out. The expression's two-dimensional intension captures this dependence: it can be seen as a function from centered worlds to 2-intensions, or equivalently as a function from pairs of centered worlds and worlds to truth-values. In the case of 'Hesperus', for example, the two-dimensional intension maps a centered world V to the 2-intension that picks out V's evening star (if it exists) in any worlds W. The actual 2-intension of an expression correspond to the two-dimensional intension evaluated at the actual centered world of the speaker: given that Venus is the actual world's evening star, the 2-intension of 'Hesperus' picks out Venus in all worlds. The 1-intension of an expression can be reconstructed by "diagonalizing" the two-dimensional intension: one evaluates it at a centered world W, and then evaluating the resulting 2-intension at the same world (stripped of its center). One might think of the two-dimensional intension as representing the way that an expression can be used to evaluate counterfactual worlds, depending on which world turns out to be actual.

I will return to these themes later, but for now it must be acknowledged that the situation is much more complicated than I have made things sound. A number of different two-dimensional frameworks have been introduced, and many of these given different results. A partial list of proponents of these frameworks, along with the names they give to their two-dimensional notions, includes:

Kaplan (1974; 1989): character and content

Stalnaker (1978): diagonal proposition and propositional expressed

Evans (1979): deep necessity and superficial necessity

Davies/Humberstone (1981): "fixedly actual" truth and necessary truth

Chalmers (1996): primary intension and secondary intension

Jackson (1998): A-intension and C-intension

There are many differences between these frameworks, some on the surface, and some quite deep. Surface differences include the fact that where Chalmers and Jackson speak of two sorts of intensions, Evans and Davies/Humberstone speak of two sorts of necessity, while Kaplan and Stalnaker speak of propositions. This sort of difference is mostly intertranslatable. Given a notion of necessity and a corresponding way of evaluating possibilities (as with Evans and Davies and Humberstone), one can define a corresponding sort of intension, and vice versa. Stalnaker's propositional content is just a set of possible worlds, which is equivalent to the intension of a sentence, and Kaplan's content is closely related.[*] Kaplan's and Stalnaker's first-dimensional notions are defined over contexts (which are at least closely related to centered worlds), and initially involve a two-dimensional intension: a function from contexts to 2-intensions. Stalnaker diagonalizes this function, yielding a function from contexts to truth-values, or a 1-intension. Kaplan leaves his character as a two-dimensional function from contexts to 2-intensions, but a corresponding step could straightforwardly be taken. So in all these cases, there is a similar formal structure.

*[[Kaplan's content is strictly speaking a singular proposition rather than a set of worlds, but it immediately determines a set of worlds. For our purposes, the difference between singular propositions, other structured propositions, and sets of worlds in analyzing the second dimension of content will not be crucial, so for simplicity I will speak as if the relevant second-dimensional contents are intensions. The discussion can be straightforwardly adapted to other views.]]

At a conceptual level, the frameworks have something further in common. In each case, the first-dimensional notion is put forward at least in part as a way of better capturing the cognitive or rational significance of an expression than the second dimension. And in each case, at least some sort of link between the first-dimensional notion and apriority has been claimed. In Kaplan's and Stalnaker's original publications, it is held that character and diagonal propositions closely reflect matters of apriority, at least in some cases. For Evans and Davies and Humberstone, when a statement of a certain sort is knowable a priori, it is deeply necessary, or true fixedly actually. And for Chalmers and Jackson, when a sentence is a priori, it has a necessary primary intension or 1-intension.

But these similarities mask deep underlying conceptual differences. These frameworks are defined in quite different ways, and apply to quite different items of language, yielding quite different results. Correspondingly, proponents of these frameworks differ greatly in the scope and strength of their claims. Kaplan's analysis is restricted to just a few linguistic expressions: indexicals and demonstratives. He explicitly resists an extension of his framework to other expressions, such as names and natural kind terms. Evans and Davies and Humberstone define their notions for a different narrow class of expressions: descriptive names, and perhaps (in the case of Davies and Humberstone) a few other terms, such as certain natural kind terms. Stalnaker's analysis applies in principle to any sentence; but in recent work, he has explicitly disavowed any strong connection with apriority, and has been skeptical about applications of two-dimensional semantics in that direction. By contrast, Chalmers and Jackson suggest

that their notions are defined for a very wide class of expressions, and make strong claims about the connection between these notions and apriority.[*]

*[[The current paper might be viewed in part as a defense of these strong claims, in light of doubts expressed by Stalnaker and others.]]

These differences arise from different *interpretations* of the formal two-dimensional framework. The framework of worlds and intensions, taken alone, is simply an abstract structure in need of content. Different interpretations flesh out this content in different ways. The interpretations are not necessarily incompatible, although it is possible that some are ill-defined, or rest on false presuppositions. The relations between these interpretations, however, are not well-understood.

The main project of this paper is to explore the different ways in which a two-dimensional framework can be understood. What are the fundamental concepts underlying different interpretations of the framework? How are these related? How do the differences between the foundations of these interpretation explain the differences in the scope and strength of the claims that are made for them? Which interpretations of the framework yield the strongest connections between the first dimension and the rational domain?

The central question on which I will focus is the following. Is there an interpretation of the two-dimensional framework that yields constitutive connections between meaning, meaning, and possibility? That is, is there an interpretation on which the first dimension is tied universally to the rational domain? On this way of thinking, the ideal form of the two-dimensional framework will recapture something like the neo-Fregean thesis: two terms will have the same 1-intension if and only if they are equivalent a priori. To get at this question, we can focus on the following core thesis:

Core Thesis: For any sentence S, S is a priori iff S has a necessary 1-intension.

The Core Thesis can be seen as the two-dimensionalist's version of the Fregean or neo-Fregean theses. It links the rational notion of apriority, the modal notion of necessity, and the semantic notion of intension. If the Core Thesis is true, it restores a golden triangle of connections between meaning, reason, and possibility.

If the two-dimensional framework can be understood in such a way that the Core Thesis is true, it promises an account of a broadly Fregean aspect of meaning, tied constitutively to the epistemic domain. It also promises further rewards: perhaps an account of the contents of thought on which content is tied deeply to a thought's rational role (potentially yielding an account of so-called "narrow content" and "modes of presentation" in thought), and perhaps a view of modality on which there are deep links between the rational and modal domains (potentially grounding a connection between notions of conceivability and possibility). So the key question in what follows will be: can we define 1-intensions so that the Core Thesis is true?

To anticipate, my answer will be as follows. There are two quite different ways of understanding the two-

dimensional framework: the *contextual* understanding and the *epistemic* understanding. The contextual understanding uses the first dimension to capture *context-dependence*. The epistemic understanding uses the first dimension to capture *epistemic dependence*. The contextual understanding is more familiar, but it cannot satisfy the Core Thesis. The epistemic understanding is less familiar, but it can satisfy the Core Thesis. The reason is that only on the epistemic understanding is the first dimension constitutively tied to the epistemic domain.

Within each of these general understandings of the framework, there are various possible specific interpretations. In what follows, I will first explore contextual interpretations, and then epistemic interpretations. Some of these interpretations are closely related to existing proposals, but rather than working directly with existing proposals, I will characterize these interpretations from first principles. This allows us to examine the properties of these interpretations in a clear light, free of problems of textual exegesis. Later in the paper, I will examine how existing proposals fit into this scheme.

(A note before we begin: many of the proposals I will examine assign intensions to expression *tokens*, such as utterances, rather than to linguistic expression types. So the Core Thesis should be understood as applying to sentence tokens. This requires that apriority be understood as a property of sentence tokens: to a first approximation, we can say that a sentence token is a priori when it expresses a thought that can be justified independently of experience. This also entails that the semantic values we are dealing with are not constrained to be varieties of "linguistic meaning", the sort of semantic value that is common to all tokens of an expression type. I will say more about these things later.)

2 The contextual understanding

On the contextual understanding of two-dimensional semantics, the possibilities involved in the first dimension represent possible *contexts of utterance*, and the intension involved in the first dependence represents the *context-dependence* of an expression's extension. There are many ways in which the extension of an expression can depend on the context in which it is uttered. On the contextual understanding, a 1-intension captures the way in which an expression's extension depends on its context. As we will see, this sort of context-dependence can itself be understood in a number of different ways.

To formalize this, we can start by focusing on expression *tokens*: spoken or written tokens of words, sentences, and other expressions. We can take it that any expression token has an extension. In cases where a token "aims" to have an extension but fails, as with an empty name, we can say that it has a null extension. If there are some expression tokens that do not even aim to have an extension (as perhaps with some exclamations), they are outside the scope of our discussion. A token of a sentence is an *utterance*; its extension is a truth-value.

Any expression token falls under a number of different expression *types*. A token may fall under an orthographic type (corresponding to its form), a semantic type (corresponding to its meaning), a linguistic type (corresponding to its type within a language), and various other types. Different tokens of the same expression type will often have different extensions. When two tokens of the same expression type have

different extensions, this reflects a difference in the *context* in which the tokens are embedded.

For our purposes, contexts can be modeled as centered worlds. The context in which an expression token is uttered will be a centered world containing the token. This can be modeled as a world centered on the speaker making the utterance, at the time of utterance. It is also possible to model a context by a different sort of centered world with just an expression token marked at the center, but the previous version will work for most purposes, as long as we assume that a subject makes at most one utterance at a given time.

One can now define the *contextual intension* of an expression type. This is a function from centered worlds to extensions. It is defined at worlds centered on a subject uttering a token of the expression type. At such a world, the contextual intension returns the extension of the expression token at the center.

One can also define the contextual intension of an expression token, relative to a type of which it is a token. This is also a function from centered worlds to extensions. It is defined at worlds centered on a token of the same type, and returns the extension of the token at the center. This contextual intension is the same as the contextual intension of the relevant expression type.

The first-dimensional intensions in the two-dimensional framework are often understood as contextual intensions of some sort. On this way of seeing things, a 1-intension mirrors the evaluation of certain metalinguistic subjunctive conditionals: if a token of the relevant type were uttered in the relevant context, what would its extension be? Of course, for every different way of classing expression tokens under types, there will be a different sort of contextual intension. In what follows I examine some of the relevant varieties of contextual intension.[*]

*[[The discussion in this section owes a significant debt to the work of Robert Stalnaker and Ned Block. Stalnaker has stressed the importance of constructs akin to contextual intensions in a number of writings (e.g. Stalnaker 1978, 1999); at the same time, Stalnaker and Block have both been active critics of the overextension of this framework (e.g. Stalnaker 1990, 2001; Block 1991; Block and Stalnaker 1999). Although I carve up the territory in a different way, a number of the varieties of contextual intension that I mention are touched on explicitly or implicitly by Stalnaker and Block at various points; and one of main points in criticizing contextual intensions as a quasi-Fregean aspect of meaning echoes points made by Stalnaker and by Block in criticizing the use of the two-dimensional framework in accounting for narrow content.]]

2.1 Orthographic Contextual Intensions

We can say that two tokens are tokens of the same *orthographic type* when they have the same orthography. This holds roughly when they are made up of the same letters or sounds, regardless of their meaning, and regardless of the language in which they are uttered. The exact details of what counts as the same orthography can be understood in different ways, but these differences will not matter for our purposes.

Then: the *orthographic contextual intension* of an expression token T is defined at centered worlds with a token of T's orthographic type at the center. It maps such a world to the extension of the relevant token in that world.

As an example, let S be Oscar's utterance of 'water is H_2O '. Let W_1 be Oscar's world (Earth), centered on Oscar making this utterance. Oscar's utterance is true, so S 's orthographic contextual intension is true at W_1 . Let W_2 be a universe containing Twin Earth (where everything is just as on earth except that the watery liquid is XYZ), centered on Twin Oscar uttering 'water is H_2O '. Twin Oscar's utterance is false (his word 'water' refers to XYZ), so S 's orthographic contextual intension is false at W_2 . Let W_3 be a universe containing Steel Earth, where the word 'water' refers to steel, but chemical terms are the same, centered on Steel Oscar uttering 'water is H_2O '. Steel Oscar's utterance is false, so S 's orthographic contextual intension is false at W_3 .

(The orthographic contextual intension of a sentence token is quite similar to its *diagonal proposition*, as defined by Stalnaker (1978). I will return to this matter later.)

It is clear that orthographic contextual intensions do not satisfy the Core Thesis. For *every* orthographic type, some possible token of that type expresses a falsehood. For example, there are worlds in which the string 'bachelors are unmarried' means that horses are cows. In such a centered world, the orthographic contextual intension of 'bachelors are unmarried' are false. The same goes for any sentence. So no truth has a necessary contextual intension, and in particular no a priori truth has a necessary contextual intension. So if 1-intensions are understood as orthographic contextual intensions, the Core Thesis is obviously false.

2.2 Linguistic contextual intensions

We can say that two expression tokens are tokens of the same *linguistic type* when they are tokens of the same expression in a language. This assumes that expression tokens belong to languages, and that languages involve expressions such as words, phrases, and sentences. So any two tokens of the English word 'water' share a linguistic type, as do any two utterances of the French sentence 'C'est la vie'.

Then: the *linguistic contextual intension* of an expression token T is defined at centered worlds with a token of T 's linguistic type at the center. It maps such a world to the extension of the relevant token in that world.

As before, let S be Oscar's utterance of 'water is H_2O '. If W_1 is Oscar's own centered world (Earth): S 's linguistic contextual intension is true at W_1 . If W_2 is Twin Oscar's centered world (Twin Earth): it is arguable that Twin Oscar's word 'water' is a *different* word from Oscar's word 'water'. Certainly *if* the referent of 'water' is essential to the word, as many theorists hold, then Twin Oscar's 'water' is a different word. If so, S 's linguistic contextual intension is not defined at W_2 . If W_3 is Steel Oscar's centered world (where 'water' means steel): here it is reasonably clear that Steel Oscar's 'water' is a different word that has the same orthography. If so, S 's linguistic contextual intension is not defined at W_3 . Applying this sort of reasoning, one reaches the conclusion that S 's contextual intension is true at every world in which at which it is defined, since the English word 'water' refers to H_2O in every world in which it exists, and so does the English expression ' H_2O '.

If this is right, then linguistic contextual intensions do not satisfy the core thesis. 'Water is H₂O' is a posteriori, but it seems to have a necessary contextual intension, true at every world at which it is defined. The same goes even more clearly for sentences involving names, such as 'Cicero is Tully'. It is widely held that names have their referents essentially; if so, the linguistic contextual intensions of true identities of this sort will be true at all worlds at which they are defined. As such, linguistic contextual intensions do not behave at all like Fregean senses. If 1-intensions are understood as linguistic contextual intensions, the Core Thesis is false.

There are some expressions for which linguistic contextual intensions behave more like Fregean senses. One such is 'I': setting certain odd cases aside, any token of the English word 'I' picks out the utterer of that token. So the linguistic contextual intension of 'I' picks out the speaker at the center of any centered world at which it is defined. In this way, it behaves much as we earlier suggested the 1-intension of 'I' should behave. Something similar applies to other indexicals, such as 'today', and to some broadly descriptive terms, such as 'philosopher'. It is in the case of names and natural kind terms that the fit seems to be worst.

(The linguistic contextual intension of an expression is in some respects like its *character*, as defined by Kaplan, although there are some differences, as we will see.)

2.3 Semantic contextual intensions

We can say that two expression tokens are tokens of the same *semantic type* when they have the same semantic value. An expression token's semantic value is its meaning or content, or some aspect of its meaning or content. There are many different ways of assigning semantic values to expression tokens, so there are correspondingly many different ways of classing expression tokens under semantic types.

Then: the *semantic contextual intension* of an expression token T is defined at centered worlds with a token of T's linguistic type at the center. It maps such a world to the extension of the relevant token in that world.

As before, let S be Oscar's utterance of 'water is H₂O'. If W₁ is Oscar's own centered world (Earth): S's semantic contextual intension is true at W₁. If W₂ is Twin Oscar's centered world (Twin Earth): at least on many ways of assigning semantic value, Twin Oscar's term 'water' has a different semantic value from Oscar's, so S's semantic contextual intension (for this sort of semantic type) is undefined at W₂. If W₃ is Steel Oscar's centered world, then Steel Oscar's term 'water' clearly has a different semantic value from Oscar's, so S's semantic contextual intension is undefined at W₃. If W₄ is a world centered on 'French Oscar', a counterpart of Oscar who speaks French and is uttering 'eau est H₂O': then it is plausible that this utterance has the same semantic value as Oscar's, so S's semantic contextual intension is defined at W₁ and is true there.

Of course the behavior of a semantic contextual intension will depend on our choice of semantic value. For example, if we stipulate that the relevant semantic value of an expression is its *extension*, then any

two co-extensive expressions will have the same semantic contextual intension, and there is no chance that the Core Thesis will be true. There are two choices of semantic value that are a little more interesting, however.

We might stipulate that the relevant semantic value of an expression is its *linguistic meaning*: roughly, the aspect of meaning that is common to all tokens of the expression's linguistic type. If we do this, then an expression's semantic contextual intension will be an extension of its linguistic contextual intension to a broader space of worlds. At worlds centered on a token of the same linguistic type, the intensions will give the same results. But the semantic contextual intensions will also be defined at other worlds, centered on synonyms and translations of the original expression. Nevertheless, if 1-intensions are understood as these semantic contextual intensions, the Core Thesis will be false for the same reasons as in the case of linguistic contextual intension. For example, if the extension of 'water' is essential to the word, then it is part of the word's linguistic meaning. So the semantic contextual of 'water is H₂O' will be true at every world where it is defined, and the Core Thesis is false.

Alternatively, we might stipulate that the relevant semantic value of an expression token is its *Fregean or descriptive content*, corresponding roughly to the expression's cognitive significance for the subject. On this reading, the Core Thesis may be more plausible. For example, one might argue that Oscar's and Twin Oscar's terms 'water' have the same descriptive content. If so, then the semantic contextual intension of Oscar's utterance 'water is H₂O' is defined at W₂ and is false there. On the other hand, Steel Oscar's term 'water' plausibly has a different descriptive content, so the semantic contextual intension of Oscar's utterance is not defined at W₃.

Understood this way, semantic contextual intensions behave as we might expect a Fregean 1-intension to behave, at least to some extent. One can argue that when a statement is a priori, any possible statement with the same descriptive content will be a priori and so will be true, so that the expressions semantic contextual intension will be necessary, as the Core Thesis requires. Corresponding, one might suggest that when a statement is not a priori, then there will be possible statements with the same descriptive content that are false, so that the statement's semantic contextual intension will not be necessary, as the Core Thesis requires.

I will argue shortly that this is not quite right. But even if it were right, it is clear that this sort of 1-intension cannot underwrite the full ambitions of the Fregean two-dimensionalists. The Fregean two-dimensionalist, as sketched previously, intends to use the two-dimensional framework to *ground* an aspect of meaning that is constitutively tied to meaning. But semantic contextual intensions as defined here *presuppose* such a Fregean semantic value, and so cannot independently ground such an account. If this is the best a two-dimensionalist can do, then if someone is independently doubtful about a Fregean aspect of meaning, two-dimensionalism cannot help. At best, two-dimensionalism will be a helpful tool in analyzing such a notion of meaning, given an independent grounding for the notion.[*]

*[[This sort of point is made quite clearly, in the context of discussing narrow content, by Stalnaker 1991, Block 1991, and Block and Stalnaker 1999.]]

2.4 A Further Problem

We have seen that orthographic contextual intensions are very far from satisfying the Core Thesis, while linguistic contextual intensions are closer at least in some cases, and some sort of semantic contextual intensions may be closer still. But there is a further problem that arises for any sort of linguistic or semantic contextual intension, suggesting that no such contextual intension can satisfy the Core Thesis.

Let *S* be a token of 'language exists'. Then *S* is true. Furthermore, any utterance of the linguistic item 'language exists' is true. Any utterance that *means* the same thing as 'language exists' is true. So it seems that *S* will have a necessary linguistic contextual intension, and a necessary semantic contextual intension, under any reasonable way of classifying linguistic and semantic types. But *S* is clearly a posteriori: it expresses empirical knowledge of the world, which could not be justified independently of experience. So *S* is a counterexample to the Core Thesis. So the Core Thesis is false for any sort of semantic or linguistic contextual intension.

The same goes for a number of other sentences. If *S*1 is 'words exist', then any utterance of the same expression or with the same meaning will be true. So *S*1 has a necessary linguistic and contextual intension, despite being a posteriori. If *S*2 is 'I exist', then any utterance of the same expression with the same meaning will be true, so *S*2 has a necessary linguistic and semantic contextual intension. But (somewhat controversially) *S*2 is a posteriori, justifiable only on the basis of experience. If *S*3 is 'I am uttering now', then any utterance of the same expression or with the same meaning will be true. *S*3 is clearly a posteriori, but has a necessary linguistic and semantic contextual intension.

All these cases are counterexamples to the Core Thesis. All of them are a posteriori and cognitively significant, and many of them seem to be as cognitively significant as paradigmatic expressions of empirical knowledge. But all have necessary semantic and linguistic contextual intensions. So the Core Thesis is false for all such intensions.

The trouble is that apriority and "truth whenever uttered" are fundamentally different notions. The first builds in an epistemic or rational element, but the second builds in no such element. The second build in a metalinguistic element, but the first builds in no such element. It is possible to understand the second in a way that makes it coincide with the first in many cases, in effect by building in an epistemic element into the individuation of the relevant linguistic types. But it is impossible to do so in all such cases, since the second has an ineliminable metalinguistic element that goes beyond the epistemic domain.

I think the moral is that to satisfy the Core Thesis, we must understand the two-dimensional framework in a quite different, non-contextual way. But before doing so, I will more briefly examine some further ways in which one might define a contextual intension.

2.5 Hybrid contextual intensions

Given orthographic, linguistic, and semantic types for expression tokens, it is possible to define *hybrid*

types corresponding two conjunctions of two or more of these types. One can then define corresponding *hybrid contextual intensions*.

For example, one might say that two expressions share the same *orthographic/semantic type* when they share the same orthographic type and the same semantic type. One can then define the *orthographic/semantic contextual intension* of an expression as the function that maps a world centered on a token of the appropriate orthographic/semantic type to the extension of that token.

Hybrid contextual intensions may be useful for some purposes, but it is clear that they will not satisfy the Core Thesis any better than non-hybrid contextual intensions. So I will set them aside here.

2.6 Token-reflexive contextual intensions

It is possible to define a slightly different sort of contextual intension for an expression token by focusing not on the types that the token falls under, but on the token itself. Let us assume that an expression tokens are not tied to their context essentially: a given token might have been uttered in another context. Then we can say that the *token-reflexive* contextual intension of an expression token T is a function that maps a centered world containing T to the extension of T in that world.

The precise behavior of a token-reflexive contextual intension will depend on what properties an expression token has necessarily. It is plausible that such a token has any properties necessarily, it has its orthographic properties necessarily. If so, its token-reflexive contextual intension will be a restriction of its orthographic contextual intension, obtained by eliminating worlds centered on a *different* token of the same orthographic type. One might also hold that a token has some semantic value necessarily, or that it has its linguistic type necessarily. If so, its token-reflexive contextual intension will be a restriction of its semantic or linguistic contextual intension. If an expression has more than one of these things necessarily, its token-reflexive contextual intension will be a restriction of a hybrid contextual intension. If it has further properties necessarily (e.g. its speaker), it will be a further restriction of the relevant contextual intension.

It is not obvious how to decide exactly which properties an expression token has necessarily. But however we do this, it is clear that token-reflexive contextual intensions cannot satisfy the Core Thesis. The counterexamples discussed above, such as 'language exists', will apply equally to token-reflexive contextual intensions. Furthermore: insofar as tokens have any properties necessarily, one can likely construct sentence tokens attributing these properties that are true whenever uttered, but not a priori (e.g. 'This token has four words'; 'David Chalmers is speaking now'). And insofar as tokens have few properties necessarily, one can likely construct sentences that are a priori but that are not true whenever uttered (e.g. 'all bachelors are unmarried'). So if 1-intensions are understood as token-reflexive contextual intensions, the Core Thesis is false.

2.7 Extended contextual intensions

In an attempt to get around the problems posed by sentences such as "language exists", one might attempt to construct contextual intensions that are defined at centered worlds that do not contain a token of the relevant expression type. The most obvious way to do this is to appeal to certain counterfactual conditionals. Let us say that the *extended contextual intension* is defined at any centered world, independently of whether a token of the type is present there. At a given centered world, the extended contextual intension returns what the extension of a token of that type *would be*, if it *were* uttered at the center of that world.

One can then say that the extended contextual intension of an expression token (relative to a type) is maps a centered world to what the extension of a token of that type would be, if it were uttered at the center of the world. So in principle, one might have extended linguistic contextual intensions, extended semantic contextual intensions, and so on. One could define an extended token-reflexive contextual intension in an analogous way.

An obvious problem here is that in many cases, it is unclear how to evaluate the counterfactual. It may be reasonably straightforward in some cases, such as 'I am a philosopher': true just when an utterance of 'I am a philosopher' by the subject at the center would be true, so true just when the person at the center is a philosopher. But how is one to evaluate what a token of 'water' would refer to if it were used in a world where there is no liquid, and in which nobody speaks a language? How does one evaluate whether an utterance of 'I am speaking loudly' would be true if it were uttered, in a world where the subject at the center is not in fact speaking? In some cases, it seems impossible for a token of the relevant type to be uttered in the relevant context. In other cases, it may be possible, but it is possible in many different ways, yielding many different results. So the truth of the relevant counterfactuals seems to be underdetermined, and an expression's extended contextual intensions seems to be ill-defined.[*]

*[[A point of this sort is made by Stalnaker 1990.]]

Another problem: even if extended contextual intensions behave coherently, they give results that are different from what we need. For example, let S = 'language exists'. S is a posteriori, so the Core Thesis requires that its 1-intension be false at some worlds. Intuitively, it is desirable that S 's 1-intension be false at a language-free world. Let W be such a language-free centered world. To evaluate S 's contextual intension at W , we ask: if S were uttered at the center of W , what would its extension be? It is clear that if S *were* uttered in W , it would be true. So S 's extended contextual intension is true at W , and indeed is true at all worlds. So the Core Thesis is still false for extended contextual intensions.

To get anything like the result that is needed, we would need to evaluate S 's extension in W *without* S being present in W . But it is very hard to do that on the contextual model. On the contextual understanding, 1-intensions are derivative on facts about the extensions of various possible tokens, as uttered in various possible contexts. It seems clear that on such an understanding, the 1-intension of a sentence such as 'language exists' will never be false.

I think that the idea of an extended contextual intension is getting at something important: that we need to be able to evaluate an expression's 1-intension in centered worlds that do not contain a token of the

expression. But this is the wrong way to achieve the goal. To do this properly, I think we need to go beyond the contextual understanding of 1-intensions.

2.8 Cognitive contextual intensions

One might suggest that to capture a token's cognitive significance, we should not focus on a token's broadly *linguistic* properties, such as its orthography, its semantic value, and its language. Instead, we need to focus on its *cognitive* properties, which correspond to mental features of the subject that produces the token. Some such features include: the concept or belief that the token expresses; the cognitive role associated with the token; and the intentions associated with the token. Assuming that we have a way of individuating the mental types in question, we can then classify expression tokens under corresponding cognitive types.

For a given scheme of cognitive typing, one can then define the *cognitive contextual intension* of an expression token as the intension that maps a world centered on a token of the same cognitive type to the extension of that token. In the three cases above: a *conceptual* contextual intension will be defined at worlds centered on a token expressing the same concept or belief; a *cognitive-role* contextual intension will be defined at worlds centered on a token associated with the same cognitive role; and an *intention-based* contextual intension will be defined at worlds centered on a token associated with the same intentions.

Assuming that one can make sense of the relevant typing, there is a natural extension of this idea. One could define a sort of *extended* cognitive contextual intension, defined at worlds that do not contain the token at all, but merely contain the relevant mental feature. For example, the extended conceptual contextual intension will be defined at any world that contain the relevant *concept* at its center, irrespective of whether it contains any token, and will return the extension of the concept. (This assumes that concepts have extensions, which seems reasonable enough.) The extended cognitive-role contextual intension might be defined at any world centered on a concept that plays the relevant cognitive role, returning the concept's extension; and the extended intention-based contextual intension will be defined at any world centered on a concept is associated with the same intentions.

This sort of intension has some promise of dealing with the central problems raised so far. In the case of 'language exists': one can make a case that the extended conceptual contextual intension of this expression is *false* at some centered worlds: those in which a subject has the relevant concepts and the relevant thought, but in which there is no language. So the intension is not necessary, reflecting the aposteriority of the sentence. The same goes for 'words exist', and for 'I am uttering now'. By allowing intensions to be evaluated without relying on language, the metalinguistic element of contextual intensions has been reduced or eliminated.

Still, analogous problems arise. 'I am thinking now' will plausibly have a necessary conceptual contextual intension, but it is plausibly a posteriori: the thought itself is justified only by experience, albeit by introspective experience. The same goes for 'I exist'. And the same will apply to specific attributions of

mental features: a thought such as 'I have the concept *concept*' will be true whenever it is thought, but it is not justifiable a priori. Something similar applies to thoughts attributing certain cognitive roles or certain intentions. So even here, some a posteriori sentences and thoughts will have a necessary 1-intension.

As for the other main sort of problem discussed so far, that associated with 'water is H₂O': a proponent might hold that although Oscar and Twin Oscar do not have the same *word* 'water', their words express the same *concept*, at least under one reasonable way of individuating concept types. If so, then the conceptual contextual intension of Oscar's token 'water is H₂O' will be false at the world centered on Twin Oscar, as the Fregean conception requires. At the same time, it might be undefined at the world centered on Steel Oscar (since he seems to have a different concept), as required.

It is controversial, however, whether concepts (or roles or intensions) can be individuated in a way that yields these results. Many theorists hold that even a token concept expressed by 'water' has its extension essentially, and that all concepts of the same type have the same extension. If so, then a statement such as 'water is H₂O' will have a necessary intension. They might concede that concepts or thoughts can also be individuated syntactically or formally; but on this way of doing things, 'all bachelors are unmarried' will have a contingent intension. So either way, the Core Thesis is false.

One might argue that there is an intermediate way of individuating concept types that yields the right results. But many will deny this. It might be objected that this requires individuating concepts by their *narrow content* (that aspect of their content that is in their head), and it is highly controversial whether narrow content exists. Some think that the two-dimensional framework can be used to give an account of narrow content; but in this context, it seems illegitimate for the framework to presuppose narrow content. This is a precise analog of the problem that arose for the Fregean version of semantic contextual intensions above.

I think that the situation here is not entirely clear. One could argue with some plausibility that there is an *intuitive* sense in which Oscar and Twin Oscar have the same concept, where there is no corresponding intuitive sense that they have the same word. If so, one could appeal to this intuitive sort of concept individuation to ground some sort of conceptual contextual intension here. One might arguably be able to do the same sort of thing with cognitive roles, or intentions. But the intuitions in questions are likely to be disputed by many, so this approach will be at best weakly-grounded, unless one can give some sort of independent account of the relevant concept types.

On my view, (extended) cognitive contextual intensions are the sort of contextual intensions that are closest to satisfying the Core Thesis. But ultimately, the central problems arise for them too. So we still need an independent account of the relevant intensions.

(There are numerous other possible notions that resemble cognitive contextual intensions in carrying features of the subject across worlds: an *evidential* contextual intension, requiring sameness of evidence; a *fixing* contextual intension, requiring sameness of reference-fixing procedures or intentions; a *physical* contextual intension, requiring that subjects be physical duplicates; *functional*, *phenomenal*, *physical-phenomenal* contextual intensions, which require that subjects be functional, phenomenal, and physical-

phenomenal duplicates; and so on. It is not hard to see that all of these suggestions are subject to versions of the problems mentioned above.)

2.9 Summary

>From what we have seen, it seems that contextual intensions cannot support the core thesis. Two central problems have arisen repeatedly. First, by building in a token of the relevant mental or linguistic type into the world of evaluation, the constitutive connection with the a priori is lost. Second, for a contextual intension to behave in a quasi-Fregean manner, we need to antecedently classify tokens under some sort of quasi-Fregean type, so that the framework cannot independently ground quasi-Fregean notions, as was originally hoped.

Contextual intensions may still be useful for many purposes. But they do not yield any restoration of the golden triangle, and in particular they do not deliver a notion of meaning that is deeply tied to reason. The fundamental problem is that although some contextual intensions yield a reasonably strong *correlation* with the epistemic domain, none are *constitutively* connected to the epistemic domain. To restore the connection between meaning and reason, we need to approach the two-dimensional framework in epistemic terms.

3 The epistemic understanding

3.1 Epistemic dependence

On the epistemic understanding of two-dimensional semantics, the possibilities involved in the first dimension represent *epistemic possibilities*, and the intensions involved in the first dimension represents the *epistemic dependence* of the extension of our expressions on the state of the world.

There are two central ideas here. The first is the idea that there are many ways the world might turn out to be, and that there is a corresponding space of epistemic possibilities. The second is that once we know how the world has turned out, or once we know which epistemic possibility is actual, we are in a position to determine to extensions of our terms. This yields a sort of function from epistemic possibilities to extensions, or an *epistemic intension*.

Taking the first idea first: there are many ways the world might be, for all we know. And there are even more ways the world might be, for all we know a priori. The oceans might contain H₂O or they might contain XYZ; the evening star might be identical to the morning star or it might not. These ways the world might be correspond to epistemically possible hypotheses, in a broad sense. Let us say that a claim is *epistemically possible* (in the broad sense) when it is not ruled out a priori. Then it is epistemically possible that water is H₂O, and it is epistemically possible that water is XYZ. It is epistemically possible that Hesperus is Phosphorus, and epistemically possible that Hesperus is not Phosphorus.

Just as one can think of metaphysically possible hypotheses as corresponding to an overarching space of metaphysical possibilities, one can think of epistemically possible hypotheses as corresponding to an overarching space of epistemic possibilities. Some possibilities in the space of metaphysical possibilities are maximally specific: these can be thought of as *maximal metaphysical possibilities*, or as they are often known, possible worlds. In a similar way, some possibilities in the space of epistemic possibilities are maximally specific: these can be thought of as *maximal epistemic possibilities*, or as I will call them, *scenarios*.[*]

*[[The term 'scenario' has been used technically by other philosophers for various purposes, notably by Peacocke (1992). My use differs from Peacocke's, but at least there is a relationship, in that Peacocke's scenarios can be seen as akin to perceptual hypotheses. Peacocke's scenarios correspond to specific *spatial* ways things might be in the vicinity of a perceiver, as presented by a perceptual state. A scenario in Peacocke's sense corresponds to a group of scenarios in my sense such that each verifies certain spatial/indexical claims: in my terms, we might call such an entity a spatial scenario type.]]

A scenario corresponds, intuitively, to a maximally specific way the world might be, for all one can know a priori. Scenarios stand to epistemic possibility as possible worlds stand to metaphysical possibility. Indeed, it is natural to think of a scenario as a sort of possible world, or better, as a *centered* possible world. There are some complications here, but to sketch some basic ideas, we can think of scenarios intuitively in such terms for now.

For any scenario, it is epistemically possible that the scenario is actual. Intuitively speaking, for any world W, it is epistemically possible that W is actual. And for any centered world W, it is epistemically possible that W is actual. Here the center represents a hypothesis about my own location within the world. In entertaining the hypothesis that W is actual, I entertain the hypothesis that the actual world is qualitatively just like W, that I am the subject at the center of W, and that now is the time at the center of W.

For example, let the XYZ-world be a specific centered "Twin Earth" world, in which the subject at the center is surrounded by XYZ in the oceans and lakes. Then no amount of a priori reasoning can rule out the hypothesis that the XYZ-world is my actual world: i.e., that I am in fact living in such a world, where the liquid in the oceans and lakes around me is XYZ. So the XYZ-world represents a highly specific epistemic possibility.

When we think of a world as an epistemic possibility in this way, we are *considering it as actual*. On the epistemic understanding, to consider a world W as actual is to consider the hypothesis that W is one's own world. Of course when one considers such a hypothesis, one does not simply consider W "neat": rather, one must consider W under a description. In effect, one is considering the hypothesis that D is the case, where D corresponds to a *canonical description* of W. We might think of D intuitively as a neutral qualitative description of W, combined with relevant indexical claims about one's own location at the center. I will return to this matter later.

The second key point is that there is a strong epistemic dependence of an expression's extension on the state of the world. If we come to believe that the world has a certain character, we are in a position to

conclude that the expression has a certain extension. And if we were to learn that the world has a different character, we would be in a position to conclude the expression has a different extension. That is: we are in a position to come to know the extension of an expression, depending on which epistemic possibility turns out to be actual.

If we take the case of 'water is H₂O': we can say that given that the world turns out as it actually has, with H₂O in the oceans and lakes, then it turns out that water is H₂O. So if the H₂O-world is actual, water is H₂O. But if we were to discover that the oceans and lakes in the actual world contained XYZ, we would judge that water is XYZ. And even now, we can judge: *if* it turns out that the liquid in the oceans and lakes is XYZ, it will turn out that water is XYZ. Or we can simply say: if the XYZ-world is actual, then water is XYZ.

The same goes more generally. If W₁ is a specific scenario in which the morning and evening star are the same, and W₂ is a scenario in which the morning and evening star are different, then we can say: if W₁ is actual, then Hesperus is Phosphorus; if W₂ is actual, then Hesperus is not Phosphorus. The same goes, in principle, for a very wide range of scenarios and statements. Given a statement S, and given enough information about an epistemically possible state of the world, we are in a position to judge whether, *if* that state of the world obtains, S is the case.

All this is reflected in the way we use language to describe and evaluate epistemic possibilities. It is epistemically possible that water is XYZ. It is also epistemically possible that the XYZ-world is actual. And intuitively speaking, the epistemic possibility that the XYZ-world is actual is an *instance* of the epistemic possibility that water is XYZ. We can say as above: *if* the XYZ-world turns out to be actual, it will turn out that water is XYZ. We might also use a straightforward indicative conditional: if the XYZ-world is actual, then water is XYZ. Or we can use the Ramsey test, commonly used to evaluate indicative conditionals: if I hypothetically accept that the XYZ-world is actual, I hypothetically conclude that water is XYZ.

We can put all this by saying that the XYZ-world *verifies* 'water is XYZ', where verification is a way of expressing the intuitive relation between scenarios and sentences described above.[*] Intuitively, a scenario W verifies a sentence S when the epistemic possibility that W is actual is an instance of the epistemic possibility that S is the case; or when we judge that if W turns out to be actual, it will turn out that S is the case; or if the indicative conditional 'if W is actual, then S is the case' is rationally assertible, or if hypothetically accepting that W is actual leads to hypothetically concluding that S is the case. We can also say that when W verifies S, W makes S true when it is *considered as actual*. Verification captures the way that we use language to describe and evaluate epistemic possibilities.

*[[The term 'verify' is used for a related idea in Evans 1979. See also Yablo 1999.]]

This dependence can be represented by the *epistemic intension* of a sentence S. This is a function from scenarios to truth-values. If a scenario W verifies S, then S's epistemic intension is true at W; if W verifies ~S, then S's epistemic intension is false at W; otherwise, S's epistemic intension is indeterminate at W. So the epistemic intension of 'water is XYZ' is true at the XYZ-world.

Epistemic intensions resemble contextual intensions in some superficial respects, but they are fundamentally quite different. The central difference, as we will see, is that epistemic intensions are defined in epistemic terms. From what we have seen so far, epistemic intensions behave at least somewhat as one would like a quasi-Fregean epistemic intension to behave. But to investigate this matter, we must define the relevant notions more precisely.

3.2 Epistemic intensions

The intuitive picture of the epistemic understanding above can be regarded as capturing what is essential to an epistemic understanding *per se*. This intuitive picture suffices for some purposes, but nevertheless it calls for a fine-grained analysis to make clear how the picture works in detail. What follows is one way to flesh out the details. Not all of these fine-grained details should be seen as constitutive of an epistemic understanding, but I think they provide a natural way of fleshing out such an understanding.[*]

*[[Note that some of these details are necessarily complex, and some readers may prefer to skim the remainder of this section or skip ahead to section 4 on a first reading. Some other papers cover some of this material in more depth: notably, Chalmers (2002e), which covers the issues in 3.4 in more detail; Chalmers and Jackson (forthcoming), which is especially relevant to the issues in 3.6; and Chalmers (2002p) and (2002s), which discuss a number of aspects of these issues that are not discussed here.]]

The epistemic intension of a sentence token is a function from a space of scenarios to the set of truth-values, defined as follows:

The epistemic intension of a sentence token S is true at a scenario W iff D epistemically necessitates S , where D is a canonical description of W .

When the conditions specified here obtain, we can also say that W *verifies* S . The epistemic intension of S will be false at W when W verifies $\sim S$, and it will be indeterminate at W when W verifies neither S nor $\sim S$.

It remains to clarify three notions: the notion of a scenario, that of a canonical description, and that of epistemic necessitation. I investigate each of these in what follows.

3.3 Deep Epistemic Possibility and Epistemic Necessitation

First, we need to say more about epistemic possibility. The epistemic understanding of two-dimensional semantics is grounded in a notion of *deep epistemic possibility*, or equivalently, of *deep epistemic necessity*. In the ordinary sense, we say that S is epistemically possible roughly when S may be the case for all we know, and that S is epistemically necessary roughly when we are in a position to know that S is the case. A notion of deep epistemic necessity goes beyond this sort of dependence on the shifting state of an individual's knowledge, to capture some sort of rational *must*: a statement is deeply epistemically

necessary when in some sense, it rationally must be true.

Such a notion can be understood in various ways, but for our purposes there is a natural candidate. We can say that *S* is deeply epistemically necessary when it is *a priori*: that is, when the thought expressed by *S* can be justified independently of experience, yielding a priori knowledge (more on this below). Then *S* is deeply epistemically possible when the negation of *S* is not epistemically necessary: that is, when the thought that *S* expresses cannot be ruled out a priori. Henceforth, I will usually drop the modifiers "deep" and "deeply", and speak simply of epistemic possibility and necessity.

In this sense, 'water is XYZ' is epistemically possible: one cannot know a priori that water is not XYZ. In the same way, 'Hesperus is not Phosphorus' is epistemically possible, as is 'I am not a philosopher'. On the other hand, 'Some bachelors are married' is not epistemically possible, and 'All bachelors are married' is epistemically necessary. Similarly, one can argue that 'Hesperus has never been visible in the evening sky' is epistemically impossible, and its negation is epistemically necessary.

A claim is deeply epistemically possible, intuitively speaking, when it expresses a rationally coherent hypothesis about the actual world. The standards of rational coherence here are in one sense weaker than usual: if a hypothesis conflicts with empirical knowledge, it may still be deeply epistemically possible. The standards are in another sense stronger than usual: if a hypothesis can be ruled out only by a great amount of a priori reasoning, it is nevertheless deeply epistemically impossible. It is possible to define notions of possibility that meet different standards, but the current standards are best for our current purposes.

The epistemic necessity operator applies to both sentence types and sentence tokens. We require this as the sentences *S* whose epistemic intensions we are defining are tokens, and it is possible for two sentence tokens of the same linguistic type to have different epistemic properties (for the reasons, see section 3.8). The canonical descriptions *D* of scenarios, on the other hand, are sentence types, using expressions whose epistemic properties are fixed by the language. We also need an epistemic necessitation operator between sentence types of this sort and sentence tokens.

An epistemic necessity operator of this sort can be seen as a primitive of the system I am developing. On the picture where epistemic necessity corresponds to apriority, we can characterize its properties intuitively as follows. Let say that *thoughts* are the sort of occurrent propositional attitudes expressed by assertive sentences. Then a sentence token *S* is epistemically necessary when the thought expressed by *S* can be justified independently of experience, yielding a priori knowledge. A sentence type *D* is a priori when it is possible for a token of *S* to be epistemically necessary. A sentence type *D* epistemically necessitates a sentence token *S* when a material conditional '*D* -> *S*' is epistemically necessary, where this is understood as a possible token material conditional whose constituent token of *S* expresses the same thought as the original token. I will elaborate and refine these characterizations in section 3.8, but this understanding will suffice for present purposes.

We can now say that a scenario *W* verifies a sentence token *S* when *W*'s canonical description *D* epistemically necessitates *S*. This will hold when a material conditional '*D* -> *S*' is a priori, where the

conditional is understood as above. On this model, intuitively, a scenario W verifies a sentence S when one could in principle rule out a priori the hypothesis that W is actual but S is not the case.

This definition works naturally with the characterizations we will give of scenarios and of canonical descriptions, but it should be noted that this is not the only possible definition. There are various ways in which an epistemic framework might characterize the required relationship between D and S in other terms. One might appeal to the intuitive heuristics described earlier, for example. First, one could say that W verifies S when the epistemic possibility that D is the case is an instance of the epistemic possibility that S is the case. Second, one could say that W verifies S according to whether the indicative conditional "if D is the case, then S is the case" is intuitively correct. Third, one could appeal to the Ramsey test. Fourth, one might appeal to an underlying epistemic necessity operator other than apriority.

So the epistemic understanding of the two-dimensional framework here is not entirely beholden to the notion of apriority. Even if one rejects apriority, or if one rejects the application of apriority in this context, one should not reject the epistemic understanding. It is simply a *prima facie* datum that there is an epistemic dependence relations between epistemic possibilities and sentences, of the sort that was intuitively characterized earlier. One who rejects apriority will simply need to capture this dependence in other ways. Still, my own view is that the understanding in terms of apriority runs the deepest.

We can here note a fundamental difference between all of these sorts of epistemic evaluation and contextual evaluation. To evaluate a sentence S in a scenario W , there is no requirement that W contain a token of S . Even if W contains such a token, in general it will have no special role to play. All that matters is the first-order epistemic relation between D and S , not whether D says something metalinguistic about a token of S . More generally, metalinguistic facts about how a token of S would behave in certain possible circumstances play no role in defining epistemic intensions. We will see shortly how this enables us to deal straightforwardly with the problem cases for contextual intensions.

3.4 Scenarios

Scenarios are intended to stand to epistemic possibility as possible worlds stand to metaphysical possibility. This claim can be expressed by the following:

Plenitude Principle: S is epistemically possible if and only if there is a scenario that verifies S .

In effect, the Plenitude Principle says that there are enough scenarios to verify every epistemically possible claim, and that no scenario verifies an epistemically impossible claim. It is easy to see that if we understand epistemic necessity as apriority, the Plenitude Principle is equivalent to the Core Thesis. (I give it a different name to leave open the option of understanding epistemic necessity in different terms.) So the only question is whether we can understand scenarios and verification so that the Plenitude Principle is true.

Intuitively, a scenario should correspond to a maximally specific epistemically possible hypothesis, or (for short) a maximal hypothesis: a hypothesis such that if one knew that it were true, one would be in a position to know any truth by reasoning alone. (Note that talk of "hypotheses" here is intuitive; formalizations of the relevant notions will follow.) We might say that a hypothesis H1 leaves another hypothesis H2 *open* if the conjunctions of H1 with both H2 and its negation are epistemically possible. A maximal hypothesis is one that leaves no possible hypothesis open. To every scenario, there should correspond a maximal hypothesis, and vice versa.

There are two concrete ways in which we might understand scenarios. The first is the way we have already sketched: as centered possible worlds. The uncentered part of the world corresponds to a hypothesis about the objective character of one's world. The centered part is needed to handle indexical claims, such as "I am in Australia". If we are given only a full objective description of a world, numerous indexical hypotheses will be left open, so such a description does not correspond to a maximal hypothesis. Correspondingly, there are numerous epistemically possible (but incompatible) objective-indexical claims: e.g. "the world is objectively thus and I am a philosopher" and "the world is objectively thus and I am not a philosopher". We need distinct scenarios to verify these claims: hence centered worlds.

There is good reason to believe that for every centered world, there is a corresponding maximal hypothesis, at least if we describe worlds under the right sort of canonical description. (It is possible that for certain special indexical hypotheses one may need further information in the center of the world, but I will leave this matter to one side.) And one can easily make the case that an epistemically impossible sentence will be verified by no centered world (if it were so verified, it would not be epistemically impossible). So the residual question is whether there are enough centered worlds to correspond to *all* maximal hypotheses, and to verify all epistemically possible statements.

The standard Kripkean cases of statements that are epistemically possible but metaphysically impossible can be handled straightforwardly. For each such statement S, there is *some* way the world could turn out such that if things turn out that way, it will turn out that S is the case; and each of these ways the world could turn out can be seen as a centered world. In the case of 'water is XYZ', the XYZ-world is such a world; something similar applies to other cases. One might worry about how a metaphysically possible world (the XYZ-world) can verify a metaphysically impossible statement ('water is XYZ')? But two-dimensional evaluation makes this straightforward: 'water is XYZ' is true at the XYZ-world considered as actual, but false at the XYZ-world considered as counterfactual. The metaphysical impossibility of 'water is XYZ' reflects the fact that it is false at all worlds considered as counterfactual. But this is quite compatible with its being true at some worlds considered as actual.

I think that on examination, there are no clear cases of epistemically possible claims that are verified by no centered world; and in Chalmers (2002p), I have argued that there are no such cases. Still, some controversial philosophical views entail that there are such cases. For example, some theists hold that it is necessary that an omniscient being exists, but it is not a priori that an omniscient being exists. If so, "No omniscient being exists" is epistemically possible. But on this view there will be no possible world to verify it, so the Plenitude Principle for centered worlds will be false. In effect, on this view the space of

metaphysical possibilities is *smaller* in some respects than the space of epistemic possibilities.

The same goes for some other philosophical views: some views on which the laws of nature of our world are the laws of all worlds, for example; some views on which a mathematical claim (such as the Continuum Hypothesis) can be necessarily true but not knowable a priori; some views on which apparently indeterminate vague statements are true and metaphysically necessitated by underlying truths, but not epistemically necessitated by those truths; and some views where unconscious physical duplicates of conscious beings are conceivable but not metaphysically possible. If these views are correct, there will be epistemically possible claims that are not verified by any centered metaphysically possible worlds, and the Plenitude Principle for centered worlds (and the Core Thesis for epistemic intensions over centered worlds) will be false.

I have argued elsewhere that all of these views are incorrect, and rest on a false view of metaphysical possibility and necessity. A careful analysis of the roots of our modal concepts supports constitutive links between epistemic and metaphysical modal notions, and thereby supports the thesis that every epistemically possible statement is verified by some centered world. We might call this the Plenitude of Worlds Thesis, or the PW thesis. If the PW thesis is correct, the views above are incorrect, and the Core Thesis is unthreatened. Still, the PW thesis is a substantive and controversial claim about modality, and it would be nice if the defense of the current view did not have to rest on it. So it is useful to have a way of understanding scenarios that does not rest on this substantive claim about metaphysical possibilities.

The alternative is to understand scenarios in purely epistemic terms from the start. One might instead that since we want epistemic intensions to be constitutively connected to the epistemic realm, we need not invoke the metaphysical modality at all, and can instead do things in terms of the epistemic modality. There are a couple of ways one might proceed here. One could introduce the notion of a scenario (a maximal epistemic possibility) as a modal primitive, in the same way that some philosophers introduce the notion of a world (a maximal metaphysical possibility) as a modal primitive. Or one could try to *construct* scenarios directly out of materials that are already at hand.

I take the second course in Chalmers (2002e), examining a detailed construction. I do not have space to do that here, but I can give a brief idea of how one might proceed. The idea I will outline is a linguistic construction of scenarios, using linguistic items in an idealized language *L*, and a basic operator of epistemic possibility.

Let us say that a sentence *D* of *L* is *epistemically complete* when (i) *D* is epistemically possible, and (ii) there is no sentence *S* of *L* such that both *D* & *S* and *D* & ~*S* are epistemically possible. When *D* is epistemically complete, it is in effect as specific as an epistemically possible sentence can be. Let us say that *D* is *compatible* with *H* when *D* & *H* is epistemically possible, and *D* *implies* *H* when *D* & ~*H* is epistemically impossible (that is, when there is an a priori entailment from *D* to *H*). Then if *D* is epistemically incomplete, it leaves questions open: there will be *H* such that *D* is compatible with *H* but *D* does not imply *H*. If *D* is epistemically complete, *D* leaves no questions open: if *D* is compatible with *H*, *D* implies *H*. Note that *D* need not explicitly include every such hypothesis as a conjunct; they need only be implied.

Intuitively, scenarios should correspond to epistemically complete *hypotheses*, whether or not they are expressible in a language such as English. It is likely that actual languages do not have the expressive resources to express an epistemically complete hypothesis, as they are restricted to finite sentences and have a limited lexicon. So for the purposes of this construction, we need to presuppose an idealized language that can express arbitrary hypotheses. In particular, our language L should allow infinitary sentences (at least infinitary conjunctions) and should have terms that express every possible concept, or at least every concept of a certain sort. It is also important that expressions in L are *epistemically invariant*, so that there cannot be two tokens S1 and S2 of the same sentence type such that S1 is epistemically necessary and S2 is not. The exact requirements for L raise subtle issues, but we can pass over them here.

We can then focus on epistemically complete sentences of L. By the idealization, every such sentence will express a maximally specific hypothesis, and vice versa. So scenarios should correspond to epistemically complete sentences in L, although perhaps with more than one such sentence per scenario. We can say that two sentences S and T are *equivalent* when S implies T and T implies S (that is, when $S \& \sim T$ and $T \& \sim S$ are epistemically impossible). Any epistemically complete sentences in L will then fall into an equivalence class. We can now identify scenarios with equivalence classes of epistemically complete sentences in L. To anticipate the definition of verification: we can also say that a scenario verifies a sentence S (of an arbitrary language) when D implies S, where D is an epistemically complete sentence of L in the scenario's equivalence class.

Defined this way, scenarios are tailor-made to satisfy the Plenitude Principle. For any epistemically possible sentence S, one can make a straightforward case that some epistemically complete sentence of L implies S. First, because L has unlimited expressive power, some epistemically possible sentence S1 of L will imply S. Second, it is plausible that any epistemically possible sentence S1 of L is implied by some epistemically complete sentence D of L. Intuitively, to obtain D from S1, one simply conjoins arbitrary sentences that are epistemically compatible with S1 (and other conjoined sentences) until one can conjoin no more. The issue is not completely trivial, there might be endless infinitary conjunction with no maximal point, but under certain reasonable assumptions, such a sentence will exist. If so, then every epistemically possible sentence is verified by some scenario. In reverse, it is clear that any sentence verified by a scenario is epistemically possible. So the Plenitude Principle is plausibly true.

In effect, this construction formalizes the intuitive idea of a maximal hypothesis: a maximal hypothesis is equivalent to an equivalence class of epistemically complete sentences in an idealized language. We might say that where the first approach takes a *world-based* view of scenarios, on which they correspond to centered worlds, the second approach takes an *epistemic* view of scenarios, on which they correspond to maximal hypotheses.

What is the relationship between the two constructions? My own view is there is a close correspondence: every centered world corresponds to a maximal hypothesis, and every maximal hypothesis corresponds to a centered world. (Not quite one-to-one: in certain cases involving symmetrical worlds with symmetrically corresponding centers, there may be more than one centered world per maximal

hypothesis). If so, then the Plenitude Principle will plausibly be satisfied either way. But some philosophers will deny the close correspondence, because of their views about possible worlds. Such a philosopher should embrace the epistemic view of scenarios. On the epistemic view, the central theses require fewer commitments than on the world-based view; and the view is connected more purely to the epistemic realm. For this reason, one can argue that the second understanding of scenarios is more basic. Centered worlds are more familiar and are useful for various applications, however, so I will use both understandings of scenarios in what follows.

On either understanding, one scenario will be privileged with respect to any statement token as the *actualized* scenario at that token. On the world-based view, this will be the world centered on the speaker and the time of utterance. On the epistemic view, this will correspond to the maximal hypothesis that is true of the world from the speaker's perspective at the time of utterance. In general, we expect that when an expression token's epistemic intension is evaluated at the scenario that is actualized at that token, the result will be the token's extension.

3.5 Canonical descriptions

When we consider a scenario as actual, in order to evaluate an expression, we always grasp it under a description. This raises an issue. A scenario can be described in multiple ways, and it is not obvious that all such descriptions will give equivalent results. So we have to isolate a special class of *canonical descriptions* of scenarios under which they must be considered.

If we take the epistemic view of scenarios by the second construction above, the choice will be straightforward. A scenario will correspond to an equivalence class of epistemically complete sentences. Here, we can say that a canonical description of the scenario is any sentence in the corresponding equivalence class. Because all of these sentences are equivalent under implication, they will all give the same results under verification.

If we take the world-based view of scenarios, things are more complicated. Here, we require that a canonical description be a *complete neutral description* of the world. Both neutrality and completeness need explanation.

First, neutrality. To describe a world, we must choose sentences that are true of it. But will these be sentences true of the world considered as actual, or of the world considered as counterfactual? If we choose the first, there is a danger of circularity, since we need canonical descriptions to define the evaluation of worlds considered as actual. If we choose the second, there is a danger of incoherence: 'Water is H₂O' is true of the XYZ-world considered as counterfactual, but we need the XYZ-world to verify 'water is not H₂O'. So sentences such as this one need to be disbarred from canonical descriptions.

The solution is to restrict canonical descriptions to *semantically neutral* expressions. Intuitively, a semantically neutral expression is one that behaves the same whether one considers a world as actual or as counterfactual. We cannot simply *define* a semantically neutral expression in this way, since the

definition presupposes evaluation in a world considered as actual, which presupposes the notion of a canonical description. But nevertheless we have a good grasp on the notion. For example, 'water' and 'Hesperus' are not semantically neutral; but 'consciousness', 'circle', and 'cause' plausibly are. One could rely on our intuitive grasp of this notion for current purposes, or one could seek to define it.

One promising approach is to define such an expression as one that is not "Twin-Earthable". We can say that two individuals (at times) are twins if they are physical and phenomenal duplicates; we can say that two expression tokens are twins if they are produced by corresponding acts of twin speakers. Then a token is Twin-Earthable if it has a twin with a different 2-intension. This test works for many purposes. A semantically neutral term (in the intuitive sense) is never Twin-Earthable. But the reverse is not quite the case. For example, let L be an expression that functions to rigidly designate the speaker's height. Then any twin of L will have the same 2-intension (since a twin speaker will have the same height), but L is not semantically neutral. One might respond by watering down the requirements of physical and phenomenal duplication (perhaps to some sort of mental duplication), but similar cases will still arise: e.g. if M is an expression that rigidly picks out "1" if the speaker is thinking a certain sort of thought, and "0" if not, then M will be Twin-Earthable even by this sort of standard, but not semantically neutral.[*]

*[[Non-Twin-Earthability is related to Bealer's (1996) notion of semantic stability: "an expression is semantically stable iff, necessarily, in any language group in an epistemic situation qualitatively identical to ours, the expression would mean the same thing" (Bealer 1996, p. 134). It is clear that semantic stability cannot be used to characterize semantic neutrality, for the same sort of reasons as above.]]

Perhaps the most straightforward alternative to is characterize the relevant class of terms negatively. We can say that a semantically neutral term is one that involves no names, natural kind terms, or indexicals (including 'actual') either explicitly or implicitly. This rules out all terms with separate processes of reference-fixation and modal evaluation. We can also rule out all terms with context-dependent behavior, such as "heavy", in order that we can describe worlds using expression types and not just expression tokens. What is left will be purely "descriptive" terms, which are context-independent and nonrigid (at least not rigid *de jure*), which can be used to neutrally formulate the hypothesis corresponding to a world.

It would be useful to have a more positive characterization than this. The fact that we can tell in cases such as the above which terms are neutral and which are not indicates that we have an independent grasp on the notion. Perhaps the best approach is to characterize semantic wholly in terms of our modal operators of epistemic and metaphysical necessity (i.e., apriority and necessity). This might be done, but there some tricky issues. In the meantime, the optimal characterization of semantically neutral terms remains an open question.

To characterize a centered world, semantically neutral terms must be supplemented by some indexical terms, to characterize the location of a center. The best way to do this is the following. We can say that a statement is in canonical form when it has the form D & 'I am D1' & 'now is D2', where D, D1, and D2 are all semantically neutral, and D1 and D2 are identifying descriptions relative to the information in D (that is D implies 'Exactly one individual is D1' and 'Exactly one time is D2'). We can say that a *neutral description* of a centered world is a statement in canonical form such that D is true of the world, D1 is

true of the subject at the center, and D2 is true of the time at the center.

(In a few cases involving completely symmetrical worlds, there may be no identifying description of the center in semantically neutral terms. In that case, one can invoke a maximally specific description of the center instead: a description D1 such that for all D2 true of the center, D entails "everything that is D1 is D2". Here, two centered worlds that differ only in symmetrical placement of the center may yield the same canonical description. This is reasonable, as intuitively both worlds correspond to the same maximal hypothesis.)

Second, completeness. We require that a canonical description be a *complete* neutral description of a centered world. There are two possibilities here. First, we can appeal to a criterion in terms of (metaphysical) necessity. Let us say that a semantically neutral description of a world is ontologically full when it (metaphysically) necessitates all semantically neutral truths about that world, and is minimal among the class of descriptions with this property. For example, if physicalism is true, a full semantically neutral specification of fundamental physical truths will be ontologically full. Then an ontologically complete neutral description of a centered world is a neutral description where the first (non-indexical) component of the description is ontologically full.

Alternatively, we can appeal to epistemic completeness. In this sense, a complete neutral description of a centered world is simply a neutral description that is epistemically complete. This requires the claim that for any centered world, there exists an epistemically complete neutral description. This claim is nontrivial, but there are good grounds to accept it: one can argue that although non-neutral terms are modally distinctive, they do not add fundamentally new *epistemic* power to a language, so that neutral terms constitute what I call (in Chalmers 2002e) an "epistemic basis" for the space of epistemic possibilities

It is not hard to see that *if* the Plenitude of Worlds (PW) thesis from the previous section is correct, then an ontologically complete neutral description will also be an epistemically complete neutral description.[*] If so, we can then use either criterion for a canonical description. There will arguably be more explanatory power, however, in using a complete description in the ontological sense, and then allowing this description to epistemically determine all truths about a world considered as actual.

*[[The PW thesis requires the notion of verification, which requires the notion of a canonical description. For the purposes of interpreting the PW thesis above, we can assume that the canonical descriptions are required to be epistemically complete. *If* the PW thesis formulated this way is correct, ontologically complete descriptions will give the same results.]]

If the PW thesis is incorrect, then the two criteria will not coincide. An ontologically complete neutral description will not be epistemically complete, and it will leave some hypotheses unsettled (e.g. the complete physical truth about the world may leave the Continuum Hypothesis unsettled, even if it is necessarily true). If we require that canonical descriptions be ontologically complete, the epistemic intensions of these hypotheses will have an indeterminate truth-value. A consequence may be that when an expression's epistemic intension is evaluated at the actual centered world of the expression, it does not yield the expression's extension (e.g., the epistemic intension of CH may be indeterminate at the actual

world, even if CH is true). If, on the other hand, we require epistemic completeness, then the epistemic intensions of the relevant claims will have a determinate truth-value (e.g. the epistemic intension of CH will be true or false at the world according to whether CH itself is true or false there). One might do things either way, depending on one's purposes, although perhaps the second is best overall. In any case, this situation will not matter much for our purposes, since we already know that if the PW thesis is incorrect, then the Core Thesis will be false when scenarios are understood as centered worlds.

(A third alternative is to require "qualitative completeness", where this is characterized as in Chalmers (2002p) in terms of a notion of positive conceivability. This yields a notion that is usefully intermediate between epistemic completeness and ontological completeness. But I will leave this option aside here.)

It is clear that if scenarios are understood as centered worlds, the characterization of canonical descriptions is significantly more complicated than if scenarios are understood in wholly epistemic terms. This may be another point in favor of the purely epistemic understanding of scenarios.

3.6 Scrutability and nontriviality

One might have the following worry: the epistemic intension of an expression may have be well-defined, but it is *trivial*. The triviality comes from the requirement that descriptions be epistemically complete. In order for a description to be epistemically complete, it will need to specify the truth or falsity of most sentences S explicitly. For example, 'water is H₂O' will be true precisely in those scenarios that have 'water is H₂O' in their canonical description, and it will be false precisely in those scenarios that have 'water is not H₂O' in their canonical description. If this sort of thing is typical, then epistemic evaluation as defined will have an uninteresting structure.

This worry is reasonable enough, but I think that it is ultimately wrong. Epistemically complete descriptions do not need to specify the truth or falsity of most statements explicitly, and epistemic evaluation does not have a trivial structure. To see this, it is useful to focus on the *actual* world, and consider what an epistemically complete description of this world must contain. The sort of argument I give here is presented in much more depth by Chalmers and Jackson (forthcoming) and Chalmers (2002p); but here I will give the basic idea.

The second principle underlying the epistemic understanding of the two-dimensional framework was that once we know enough about the state of the world, we are in a position to know the extensions of our terms, and the truth-values of our sentences. Furthermore, we usually need not be informed about a sentence explicitly in order to know whether it is true. This might be put as follows:

Scrutability of truth: For most terms T used by a speaker, then for any truth S involving T, there exists a truth D such that D is independent of T, and such that knowing that D is the case puts the speaker in a position to know (without further empirical information, on idealized rational reflection) that S is the case.

Here, we can say that D is independent of T when D does not contain T or any close cognates. Of course this notion is somewhat vague, as is the notion of "most" above, but this does not matter for our purposes. To save breath, we can put the claim expressed above as "knowing that D... that S is the case" as "D is epistemically sufficient for S".

Take the case of 'water'. Here, we can let D be a truth specifying an appropriate amount of information about the appearance, behavior, composition, and distribution of objects and substances in one's environment, as well as information about their relationship to oneself. D need not contain the term 'water' at any point: appearance can be specified in phenomenal terms, behavior and distribution in spatiotemporal terms, composition in microphysical or chemical terms. Then D is epistemically sufficient for 'water is H₂O'. When one knows that D, one will be in a position to know all about the chemical makeup of various liquids with various superficial properties in one's environment, and will thereby be able to infer that water is H₂O. After all, this information about appearance, behavior, composition, and distribution is roughly what we need in the case of ordinary knowledge, to determine that water is H₂O. And there is no need for further empirical information to play a role here: even if we suspend all other empirical beliefs, we can know that *if* D is the case, then water is H₂O.

Similarly for terms such as 'Hesperus'. Once again, if D contains appropriate information about the appearance, behavior, composition, and distribution of various objects in the world, then D is epistemically sufficient for 'Hesperus is Venus', for 'all renates are cordates', and so on. The information in D enables one to know that the object that presents a certain appearance in the evening is the same as the object that presents a certain appearance in the morning, and so enables us to know that Hesperus is Phosphorus. Something similar applies to 'heat', 'renate', and so on. Here, the base information need not contain terms such as 'Hesperus' or 'renate', or any cognates. And no further empirical information is required: the information in the base is that is needed.

Something similar applies for terms like 'philosopher', or even names like 'Feynman'. Here, the base information D may need more than in the cases above: for example, it may need to include information about people and their mental states, and the use of certain names, and so on. But once I have this information, I will be in a position to know that Feynman was not a philosopher, even if I had no substantive knowledge of Feynman beforehand. And again, my information need not use the terms 'Feynman' or 'physicist' to do this (it might use the quite different term "'Feynman'" in order for me to track down the referent via those from who I obtained the name, but that is legitimate in this context).

This may not apply to *all* expressions. There may be primitive terms such as 'time' or 'cause' or 'conscious' such that to know whether a sentence involving these terms is true, one needs a base that includes those terms or relevant cognates that invoke them implicit. But as long as the principle applies reasonably widely, it is good enough.

By the sort of reasoning above, one can infer a slightly stronger claim. Let us say that a vocabulary is a set of terms, and that a V-truth is a truth that uses only terms in V. Then we can say: there is a relatively limited vocabulary V such that for any truth S, there is a V-truth D such that D is epistemically sufficient for S. To arrive at V, intuitively, we might simply eliminate terms one by one from the language

according to the first principle, until we cannot eliminate any further. Exactly how limited V must be is an open question, but I think the sort of reasoning above gives good ground to accept that it will involve only a small fraction of the original language. One can put the claim in a slightly stronger form:

Scrutability of Truth II: There is a relatively limited vocabulary V such that for any truth S , there is a V -truth D such that D implies S .

Here, we have moved from " D is epistemically sufficient for S " to " D implies S ": that is, that the material conditional ' $D \rightarrow S$ ' is a priori. This is a stronger but not a vastly stronger claim, given that epistemic sufficiency involved "no further empirical information". One can argue for it along much the same lines as above, suggesting that even a speaker who suspends all empirical beliefs can know that *if* D is the case, then S is the case. Chalmers and Jackson (forthcoming) argue in much more depth that this sort of conditional is a priori (for a specific choice of V). A point made there is worth noting here: this sort of a priori entailment does not require that there is an explicit definition of the terms in S using the terms in V .

(Note that even if one is skeptical about apriority, the general point about epistemic sufficiency is still plausible. So such a skeptic will be able to appeal to alternative sorts of epistemic evaluation, and will find an analogous nontriviality under those sorts of epistemic evaluation to the sorts I find here.)

It is also plausible that there is some V -truth D that implies all V -truths. Of course D may need to be an infinitary conjunction, but we may as well stipulate that V is part of our idealized language, so this is no problem. We can think of D as a conjunction of the simple V -truths about the world, or as a conjunction of all V -truths of up to a certain level of complexity. There is plausibly a level such that any more complex V -truth will be implied by this sort of conjunction. If so, it follows that D entails all truths about the world. It follows plausibly that D is epistemically complete (if D is compatible with H and $\sim H$, then all truths about the world are compatible with H and with $\sim H$, which is plausibly impossible.)

Exactly what is required for the vocabulary V and the description D is an open question. Chalmers and Jackson (forthcoming) and Chalmers (2002p) argue that a specific description D will work here: PQTI, the conjunction of microphysical and phenomenal truths with certain indexical truths and a "that's-all" truth. If this is right, then V requires only the vocabulary required for PQTI. It is possible that the vocabulary might be stripped down further, if Q is implied by P (as some physicalists hold), or if P is implied by a description in a more limited vocabulary, such as one in terms of space, time, and causal connections. But in any case, this specific claim is not required here. The only claim required is that *some* limited vocabulary V suffices for this purpose.

What goes for the actual world goes also for any epistemic possibility. There is nothing special about the actual world here. Given any class of epistemically compatible sentences in our idealized language, one can strip down the vocabulary involved in it in the same sort of way as before, until one has a limited vocabulary V such that each of the original sentences is implied by a V -sentence. It follows by similar reasoning to the above that for any scenario W , there will be a limited vocabulary V such that there is an epistemically complete V -truth that corresponds to the scenario. Of course the vocabulary may differ between scenarios. For example, there are presumably epistemically possible scenarios that involve

conceptually basic kinds that are alien to our worlds. If so, the basic vocabulary V for our world must be expanded to describe this scenario. But the resulting vocabulary will still be limited.

Further, there is good reason to accept that *most* terms of the idealized language will not be essentially required to describe any scenario. If so, there will be a relatively limited vocabulary V' such that for *any* scenario, there will be a corresponding V' -truth. We can think of V' here as an *epistemic basis*: the terms in it express a set of concepts sufficient to cover all of epistemic space. Exactly how small V' can be is again an open question, but we have good reason to believe that it need involve only a very small fraction of the terms of the original language. With V' in place, we can think of a scenario as corresponding to an equivalence class of epistemically complete V' -sentences, rather than of arbitrary epistemically complete sentences.

In any case: from all this, we can see that epistemic intensions will not be nontrivial. An epistemically complete description need not specify the status of most sentences explicitly. Most terms, such as 'water' and 'H₂O', will plausibly not be required in an epistemically basic vocabulary, so sentences involving these terms will be *nontrivially* true or false in scenarios. For all we have said here, it may be that some claims (e.g. 'there is space') are in a sense trivially true in some scenarios and trivially false in others, but this is perfectly reasonable; it is analogous to the trivial truth or falsity (in an analogous sense) of claims about ontologically fundamental properties in metaphysically possible worlds. So there will be plenty of interesting structure to epistemic intensions in general.

3.7 Subsentential epistemic intensions

So far I have defined epistemic intensions only for sentences. It is not too hard to define them for subsentential expressions, such as singular and general terms, kind terms, and predicates, but there are a few complexities. I will take it that we have already decided on independent grounds what sort of extensions these expressions should have: e.g. individuals, classes, kinds, and properties. Differences choices could be made here, but the same sort of treatment will work.

The details depend to some extent on whether we take the world-based view or the epistemic view of scenarios. The difference is that centered worlds already come populated with individuals and the like (or at least we are familiar with how to regard them as so populated), whereas maximal hypotheses do not (or at least we are less familiar with how to populate them).

If we take the world-based view of scenarios: let W be a centered world with canonical description D , and let T be a singular term. In most cases, D will imply a claim of the form ' $T=T$ ', where T is a semantically neutral singular term. If so, the epistemic intension of T picks out the referent of T in W (that is, it picks out the individual that T picks out when W is considered as counterfactual). In some symmetrical worlds, it may be that there is no such semantically neutral T , but there is a T that involves semantically neutral terms plus 'I' and 'now' (plus other basic indexicals, if any). In this case, one can replace the indexicals in T by labels for the entities at the center of the world, yielding an expression T such that the epistemic intension of T picks out the referent of T in W . If there is no such T , then the epistemic intension of T is

null in W.

One can do the same for general terms, appealing to claims of the form 'For all x, x is a T iff x is a T', and holding that the epistemic intension of T in W picks out the referent of T, for a T that is semantically neutral (perhaps plus indexicals). For kind terms, we again appeal to identities 'T is T'. For predicates, we can appeal to claims of the form 'For all x, x is T iff x is T'. This delivers extensions for the epistemic intensions straightforwardly.

If we take the epistemic view of scenarios: then we need to populate scenarios with individuals and the like. If we simply admit scenarios as a basic sort of abstract object with certain properties, one could simply stipulate that they contain individuals that can serve as the extensions of relevant expressions - much as many of those who introduce possible worlds simply stipulate something similar. But it is useful to go through an explicit construction.

Let W be a scenario with canonical description D. Let us say that two singular terms T1 and T2 are equivalent under W if D implies 'T1 is T2'. Then we can identify every equivalence class of singular terms under W with an individual in W, and hold that the epistemic intension of T in W picks out the individual corresponding to T's equivalence class in W. As for general terms: every general term G will pick out a class of individuals. One of the individuals defined above will be in G precisely when D implies 'T is a G', for some T that picks out the individual. One can do something similar for predicates and kind terms: the details will depend on the precise view one takes of properties and kinds and their relation to individuals, so I will not go into them here.

There is one worry: what if the truth of certain existentially quantified claims in a scenario requires individuals that are not the referent of any singular term? For example, there may be a predicate phi such that D implies 'There exists x such that phi(x)', and D does not imply any claim of the form 'phi(T)', where T is a singular term. Of course since D is epistemically complete, it will tell us exactly how many individuals have phi, whether some individuals with phi also have psi and some do not, and so on. It is not hard to see that this sort of case will ultimately require predicates phi (perhaps an infinitely conjunctive predicate) such that D implies that there exists more than one individual with phi, and such that for all predicates psi, D implies that these individuals are indistinguishable with respect to psi. In this case, the individuals will be indistinguishable even in our idealized language, presumably because of deep symmetries in the world. In such a case, if D implies that there are n individuals with phi, one can arbitrarily construct n individuals, perhaps as ordered pairs (phi', 1) ... (phi', n), where phi' is the phi's equivalence class, and stipulate that all fall under the extension of phi, and of other predicates and general terms as specified by the relevant D-implied universally quantified truths about individuals with phi.

In this way, we can construct the relevant classes of individuals and the like, and specify the extensions of various expressions' epistemic intensions. The construction ensures that where the extension of a complex expression is a compositional function of the extensions of its parts, then the same will be true of the extension of a complex expression relative to a scenario. For an identity (e.g. 'T1=T2'), compositionality will be ensured by the equivalence class construction. For a predication (e.g. 'T is a G', or phi(T)) this will be ensured by the appropriate construction of extensions for general terms (as above) or predicates. The

machinations two paragraphs above ensure that existential quantification will work straightforwardly, and universal quantification is guaranteed to work (if D implies $\forall x \phi(x)$, then every individual constructed above will have ϕ). Logical compositionality is guaranteed at the sentential level (if D implies S and T , D will imply $S \& T$, and so on). So the epistemic intension of a complex expression will be a compositional function of the epistemic intension of its parts.

Of course once one has engaged in this sort of construction, one need not usually bother with the details again. It is perfectly reasonable thereafter to speak of a scenario as containing individuals and the like, and to speak about terms as picking out various individuals in a scenario, quite independently of the details of the construction. On the epistemic view of scenarios, for most purposes one can think of them as abstract objects that may behave somewhat differently from possible worlds, but that have the same sort of status in our ontology.

3.8 Tokens, types, and apriority

As I have approached things, epistemic intensions have been assigned to expression tokens rather than expression types (such as linguistic types). The reason for this is straightforward. It is often the case that two tokens of the same linguistic type can have *different* epistemic intensions. This difference arises from the fact that different speakers may use the same expression so that it applies to epistemic possibilities in different ways. And this difference corresponds to the fact that different speakers may use the same term with different a priori connections.

For example, it is often the case that two speakers will use the same *name* with different a priori connections. The canonical case is that of Leverrier's use of 'Neptune', which he introduced as a name for (roughly) whatever perturbed the orbit of Uranus. For Leverrier, 'If Neptune exists, it perturbs the orbit of Uranus' was a priori. On the other hand, later speakers used the term (and still do) so that this sentence is not a priori for them: it is epistemically possible for me that Neptune does not perturb the orbit of Uranus. We can even imagine that when Leverrier's wife acquired the name, she did not acquire the association with Uranus, so that she is in no position to know the truth of this sentence a priori.

How can we characterize the epistemic intension of Leverrier's tokens of 'Neptune'? To a first approximation, we can say that in any scenario, Neptune picks out whatever perturbs the orbit of Uranus in that scenario. How can we characterize the epistemic intension for Leverrier's wife? This is a bit trickier, but we can assume that for his wife to determine the reference of 'Neptune', she would examine Leverrier's own use and see what satisfies it. So to a first approximation, his wife's epistemic intension picks out whatever Leverrier refers to as 'Neptune' in a given scenario. One can find a similar (although less stark) variation in the epistemic intensions of many names, and perhaps natural kind terms.

Something similar applies to many uses of context-dependent terms, such as 'heavy'. What I count as heavy varies with different uses of the term. In some contexts, 'My computer is heavy' may be true, and in other contexts it may be false, even though it is the same computer with the same weight. Correspondingly, the way I apply a term across epistemic possibilities will vary with these uses: if I

suppose that my computer weighs such-and-such, I may hold the utterance true in the first case but not the second.

As we have defined epistemic intensions, they are grounded in the behavior of sentences under an epistemic necessity operator. So the variation in epistemic intensions of two expressions of the same type is traceable to variations in the epistemic necessity of two type-identical sentences. In particular, it will be traceable to variations in the apriority of two type-identical sentences. And this variation is traceable to variations in the apriority of the thoughts that the two sentences express.

Here we need to say a little more about thoughts. Thoughts are understood as the occurrent propositional attitudes expressed by assertive sentences. Sentences typically express occurrent beliefs, but not always, since subjects do not always believe what is said. But when the subject does not believe the relevant content, the subject *entertains* the relevant content: a thought is an entertaining of this sort. Like beliefs, thoughts are assessable for truth. Thoughts can come to be *accepted*, yielding beliefs, and thoughts can come to be *justified*, often yielding knowledge.

On this way of approaching things, we assume a relation of expression between statements and thoughts, and we assume a notion of epistemic necessity as applied to thoughts. The latter notion might be seen as the true conceptual primitive of the approach, but on the account where epistemic necessity is tied to apriority, we can characterize it by saying: a thought is epistemically necessary when it can be justified independently of experience, yielding a priori knowledge. We can then say that a thought is epistemically possible when its negation is epistemically necessary. Two thoughts are epistemically *compatible* when their conjunction is epistemically possible. One thought *implies* another when the first is epistemically incompatible with the negation of the second. Here we assume that two thoughts of the same subject can stand in a relation of negation, and that a thought can stand in a relation of conjunction or disjunction to a set of two or more other thoughts of the same subject.

(Note that the notion of apriority involves a rational idealization. A token of a complex mathematical sentence may be a priori even if there is no way for the speaker, with her actual cognitive capacities, to justify the corresponding thought a priori. Apriority idealizes away from these cognitive capacities, by requiring merely that it would be *possible* for the thought to be justified a priori, perhaps by using cognitive capacities that the speaker does not in fact possess. This notion can be developed in detail in various ways: one such way appeals to the possibility of a mental life in which the initial thought (starting in the context of the original cognitive state) comes to be justified a priori. This sort of characterization is best regarded as a clarification of a basic intuitive notion of the epistemic necessity of a thought, rather than as a definition, in order to avoid a multiplication of primitive notions, and also in order to avoid potential problems arising from views on which certain conceivable cognitive capacities are not metaphysically possible.)

With these notions in hand, we can characterize epistemic necessity and necessitation as applied to sentences. A sentence token is epistemically necessary iff it expresses an epistemically necessary thought. A sentence type is epistemically necessary if any token of the type is epistemically necessary. If D and E are sentence types, we can say that D epistemically necessitates E when $D \& \sim E$ is epistemically

impossible. If D is a sentence type and S is a sentence token: let us say that a thought is a D -thought if it is the sort apt to be expressed by D . Then D epistemically necessitates S when a possible D -thought of the subject will imply the thought expressed by S . Equivalently, D epistemically necessitates S when if the thought expressed by S were to be disjoined with a $\sim D$ -thought, the resulting thought would be epistemically necessary.

(It is worth noting that we can use this framework to define epistemic intensions for thoughts as well as utterances. Much as above, we can say that a scenario verifies a thought when disjunction of the thought with a $\sim D$ -thought is epistemically necessary, where D is a canonical description of the scenario. This yields a notion of mental content that can be applied to beliefs, thoughts, and any propositional attitude with a mind-to-world direction of fit. One can argue that this is a sort of *narrow* content, determined by the internal state of the thinker (see Chalmers 2002c). Using the definitions above, we can see that when an utterance expresses a thought, the epistemic intension of the utterance will be identical to the epistemic intension of the thought.)

The framework involving thoughts enables us to see how two tokens of the same type can differ in apriority. If Leverrier says 'If Neptune exists, it perturbs the orbit of Uranus', his statement will presumably express a priori knowledge, and will certainly express a thought that can be justified a priori. If his wife utters the same sentence, no amount of a priori rational reflection alone could justify the thought she expresses. Similarly, if I say 'someone with 1000 hairs on their head is bald' on one occasion, it may express an a priori false thought (one whose negation is a priori justifiable), while if I say it on another occasion, it may express a thought that is not a priori false, and may be plausibly true.

In a similar way, this framework enables us to see how two tokens of the same type can have different epistemic intensions. Let S be the sentence 'Neptune is an asteroid', and let D be a canonical description of a scenario W in which the orbit of Uranus is perturbed by an asteroid and in which no-one has ever used the term 'Neptune'. (We can abstract away from complications involving the intension of 'Uranus'.) Then D epistemically necessitates Leverrier's utterance of S : a thought that D obtains would imply the thought Leverrier expresses with S . But D does not epistemically necessitate Leverrier's wife's utterance of S : a thought that D obtains would not imply the thought that his wife expresses with S . (Note that D itself will not exhibit this sort of variation, as expressions in the idealized language are required to be epistemically invariant.) So Leverrier's utterance of S is verified by W , while his wife's utterance is not. So the two utterances have different epistemic intensions.[*]

*[[The analysis in terms of thoughts also clarifies a worry about sentences such as 'Hesperus is Phosphorus'. Intuitively, such a sentence is a posteriori; but some theorists hold that such a sentence is a priori, as it expresses a trivial singular proposition that can be known a priori (e.g. by knowing that Venus is Venus). On the current definition, tokens of such a sentence are not a priori: the thought expressed by such a token clearly cannot be justified independently of experience, even if a different thought associated with the same singular proposition could be. On this understanding, the apriority of an utterance cannot be understood in terms of the apriority of an associated singular proposition. If it is understood in terms of the apriority of an associated content, the relevant content will be something more akin to an epistemic intension.]]

One might reasonably ask: in languages such as English, what sort of simple terms have epistemic

intensions that vary between speakers and occasions of use? This happens most clearly for: (i) names (such as 'Neptune' and 'Gödel'); (ii) natural kind terms (such as 'water' and 'gold'); (iii) demonstratives (such as 'that' and 'there'); and (iv) many context-dependent terms (such as 'heavy' and 'bald'). For terms like this, it is clear that an epistemic intension is not part of a term's "linguistic meaning", where this is understood as the sort of meaning that is common to all tokens of a type in a language. Instead, it is a sort of "utterance meaning" or "utterance content". Some theorists use the term "meaning" only for linguistic meaning, but this is a terminological matter; it is clear that there is a broader use of the term. What matters is that this framework yields a useful and interesting sort of semantic value in the broad sense; and this much is clear.[*]

*[[See Chalmers (2002s) for more on this matter.]]

Are there terms for which an epistemic intension is common to all tokens of a type? This is perhaps most plausible for indexicals, such as 'I' and 'today' (at least setting aside unusual uses, and any context-dependence at the boundaries). It may also hold for some descriptive terms, such as 'circle'. Most of these have some context-dependence, but this can be regimented out more straightforwardly than the epistemic variability of names and natural kind terms. Finally, it may hold for some descriptive names (e.g. 'Jack the Ripper'), at least for a certain period of their existence. For terms like these, an epistemic intension can be seen as part of their linguistic meaning.

3.9 The Second Dimension

I have concentrated almost wholly on the first dimension of the two-dimensional framework. This is because the second dimension is already well-understood. But I will say a few words about it here. It is worth examining how it can be understood in a way that is parallel to the way we have understood the first dimension.

Like the first dimension, the second dimension is founded on a certain sort of possibility and necessity. For the first dimension, this is epistemic possibility and necessity, tied to what *might be* the case. For the second dimension, this is what we might call *subjunctive* possibility and necessity, tied to what *might have been* the case.

We can say that S is subjunctively possible when it might have been the case that S. Kripke is explicit that this is the basic notion of possibility and necessity with which he is working, and almost all of his modal argument are directly grounded in intuitions about what might have been the case.

With this basic modal operator in hand, we can proceed as before. For example, one can define a *subjunctively complete* sentence parallel to the way we defined an epistemically complete sentence. One can construct equivalence classes of subjunctively complete sentences in an idealized language. One can identify these classes as *maximal metaphysical possibilities*, or as *possible worlds*. One can give possible worlds *canonical descriptions*, which will be subjunctively complete sentences in their equivalence class.

Just as one can consider a scenario as actual, by supposing that it actually obtains, one can consider a world as counterfactual, by supposing that it had obtained. That is, instead of thinking "if D is the case, then...", one thinks "if D had been the case, then..." (where D is a canonical description of a world W). For example, for a given sentence S, one can entertain and evaluate the subjunctive conditional: "if D had been the case, would S have been the case?". In some cases, the answer will intuitively be yes: in this case, we can say that W *satisfies* S. This is a distinctive sort of *counterfactual* evaluation. When W satisfies S, we can say the *subjunctive intension* of S is true at W.

The subjunctive intension of a sentence S is a function from worlds to truth-values, true at W if and only if W satisfies S. Satisfaction can be intuitively characterized as above. Formally, we can say that W satisfies S when D subjunctively necessitates S, where D is a canonical description of W. We could define subjunctive necessitation by the subjunctive conditional heuristic above. Or more formally, one might say that D subjunctively necessitates S when $D \& \sim S$ is subjunctively impossible.

With a possible world as constructed, we can construct a space of individuals much as we did with scenarios. We can then define subjunctive intensions for subsentential expressions straightforwardly.

Subjunctive intensions are defined in the first instance for expression tokens, since subjunctive necessity judgments can vary between tokens of a type. For some expression types, all tokens of the type will have the same subjunctive intension: this is arguably so for names and natural kind terms, logical and mathematical terms, and some descriptive terms (e.g. 'circle'). For other expression types, subjunctive intensions will vary between tokens of the type: this is so for indexicals (e.g. 'I') and many context-dependent terms (e.g. 'heavy'). In the first case, subjunctive intension may be an aspect of linguistic meaning; in the second case, it is not.

The basic ideas here are parallel between the two cases. The explicit construction of possible worlds and the like may seem like heavy weather; but this is only because possible world are more familiar. One might think one does not really need any such construction to legitimize the appeal to possible worlds; but if so, I would say the same applies to scenarios. Both serve as a sort of useful abstraction from a modal notion in much the same way.

There is one important difference between worlds and scenarios. We have a means of reidentifying individuals across worlds, but there is no such means of reidentifying individuals across scenarios. In the case of worlds, these claims are grounded in *de re* subjunctive intuitions of the form 'T might have been F' - read so that they are distinct in their form from *de dicto* subjunctive intuitions such as 'it might have been that T was F'. We can use these claims in conjunction with the construction above to identify certain objects in alternative possible worlds as identical to certain objects in the actual world (or alternatively, to identify objects with equivalence classes across worlds; or at least to set up counterpart relations across worlds). There is no analog of a *de re* modal intuition in the epistemic case: it can be a priori that T is F, but it makes no sense to say that T is such that it is F a priori.

(In the subjunctive case, one can also appeal here to *de dicto* subjunctive intuitions involving a privileged class of designators, i.e. names. Judgments of the form 'it might have been that N was F' where N is a

name for the relevant object, arguably give the same result for any name of the object, and if so can ground a sort of crossworld identification. In the epistemic case, there is in general no analog to this privileged class of designators: different names for an individual are not generally a priori equivalent, so come apart in different scenarios, and there is no way in general to isolate a privileged class of epistemically equivalent designators here. At best this may be possible in special cases, such as canonical designators for phenomenal states and abstract entities. A consequence is that quantified modal claims will not generally be well-defined in the epistemic case, and quantified modal logic will be largely inapplicable in this domain.)

In many cases, a term's subjunctive intension will depend on its actual extension, or on other aspects of the actual world. This is particularly clearly in the case of rigid designators such as names and indexicals. If Kripke is correct, these pick out the same individual in all possible worlds, and so pick out the term's actual extension in all possible worlds. In these cases, the subjunctive intension of a term itself depends on the character of the actual world. Here, in effect, a term's subjunctive intension depends on which epistemic possibility turns out to be actual.

One can naturally encapsulate this behavior in a *two-dimensional intension*. This can be seen as a mapping from scenarios to subjunctive intensions, or equivalently as a mapping from (scenario, world) pairs to extensions. We can say: the two-dimensional intension of a statement S is true at (V, W) if V verifies the claim that W satisfies S . If $D1$ and $D2$ are canonical descriptions of V and W , we say that the two-dimensional intension is true at (V, W) if $D1$ epistemically necessitates that $D2$ subjunctively necessitates S . A good heuristic here is "If $D1$ is the case, then if $D2$ had been the case, would S have been the case?". Formally, we can say that the two-dimensional intension is true iff $\text{nec1} (D1 \rightarrow \text{nec2} (D2 \rightarrow S))$, where nec1 and nec2 express epistemic and subjunctive necessity respectively. One can define two-dimensional intensions for subsentential expressions by an extension of this idea.

(One complication: the construction so far makes the space of possible world derive from subjunctive modal claims, so if the latter depends on the character of the actual world, the former might, too. If so, then the space of W 's may depend on V . For maximally general results, this might be useful, since some philosophical theories of modality make this claim. If every possible world can be completely specified in semantically neutral terms, the conclusion can at least be mitigated by using such a description to identify the worlds between spaces, although on some views it may still be that whether a given semantically neutral claim is possible depends on which scenario is actual.

On some views, complete semantically neutral description of possible worlds will not always exist: e.g. if there can be haecceitistic differences between worlds without qualitative differences. In this case, it is probably best to see each scenario V as being associated with a relativized space of possible worlds. Note that this worry about semantically neutral descriptions does not affect their earlier applications to centered worlds for epistemic purposes: a pair of haecceitistically different centered worlds will have the same canonical description for epistemic purposes, and will in effect function as epistemically equivalent.)

Relative to a scenario, there will be an *associated* world (in the scenario's space): intuitively, the world that will be actual if the scenario obtains. If scenarios are centered worlds, a scenario's associated world

will be the scenario stripped of its center. On the epistemic view of scenarios, we can say that (to a first approximation) W will be associated with V when $D1$ verifies $D2$, where $D1$ and $D2$ are canonical descriptions of V and W .

(One worry: on some views, it may be that there are scenarios that do not fully specify worlds, as there are epistemically complete sentences that are ontologically incomplete. For if intrinsic properties exist but can only *in principle* be specified relationally, then an epistemically complete description of a scenario may specify only relations and nothing about the nature of the intrinsic properties (the relevant claims cannot even be expressed), while nevertheless implying that there are such intrinsic properties and that they have some further nature. If this can happen, there may be no $D2$ such that $D1$ verifies $D2$. To handle this, it is probably best to say that there can be multiple worlds associated with a scenario, and that W is associated with V when $D1$ and $D2$ are epistemically compatible.)

Given the association relation between scenarios and worlds, one can define the *diagonal intension* of a sentence's two-dimensional intension. This will be a mapping from scenarios to individuals, mapping V to the value of the two-dimensional intension at (V, W) , where W is associated with V . (If there is more than one such W for the reasons above, it is not hard to see that they will all give the same results.) The diagonal intension of a sentence will straightforwardly be equivalent to its epistemic intension. One can therefore reconstruct an expression's epistemic intension from its two-dimensional intension by diagonalizing, just as one can reconstruct its subjunctive intension by holding fixed the actualized scenario.

It should be clear, however, that this diagonal construction in no sense gives the *definition* of an epistemic intension. Epistemic intensions are defined in purely epistemic terms: they are in no sense derivative on subjunctive notions. The diagonal construction is conceptually much more complex, involving subjunctive evaluation, association of worlds with scenarios. In effect, the relation is akin to that between the functions $f(x) = x^3$, $g(x, y) = x^3 + \sin(x-y)$, and $g'(x) = g(x, x)$. Here, g' is the "diagonal" of g , and is the same function as f . But it would obviously be incorrect to hold that f is fundamentally the diagonal of g , or that it is derivative on trigonometric notions. For exactly the same reasons, it is incorrect to hold that an epistemic intension is fundamentally a diagonal intension.

There is a sense in which the two-dimensional intension represents the full modal structure of an expression, capturing how it behaves under epistemic evaluation, modal evaluation, and combinations of the two. Just as an epistemic intension can be evaluated a priori, a two-dimensional intension can be evaluated a priori. A subjunctive intension cannot be, but it can be evaluated when the actualized scenario is specified.

We can think of all of these intensions as aspects of the content of a sentence token. A sentence is in no sense ambiguous for having both epistemic intensions and subjunctive intensions; rather, it has a complex semantic value. Different aspects of this semantic value will be relevant to the evaluation of the sentence in different contexts. In epistemic contexts ("it is a priori that S "; "it might turn out that S "; "if S , then T "), the epistemic intension will be most relevant. In subjunctive contexts ("it might have been that S "; "if it had been that S , it would have been that T "), the subjunctive intension will be most relevant. In

combined epistemic-subjunctive contexts, the two-dimensional intension will be most relevant. There is no need to settle the question of which of these, if any, is *the* meaning or content of an expression.

3.10 The Core Thesis

Let me summarize where things stand with respect to the Core Thesis: that S is a priori iff S has a necessary 1-intension.

If S is a priori: then for all W with canonical description D , D implies S . (If S is a priori, D implies S for *any* D .) So S is verified by all W , and has a necessary 1-intension.

If S is not a priori: then $\sim S$ is epistemically possible. Under small assumptions (see 3.2), it follows that there is an epistemically complete D such that D implies $\sim S$. On the epistemic view of scenarios, any epistemically complete sentence describes a scenario, so there is a scenario W that verifies $\sim S$. So S does not have a necessary 1-intension.

On the world-based view of scenarios, then if the PW thesis is true, any epistemically complete D describes a scenario, so S does not have a necessary 1-intension. If the PW thesis is false, this does not follow. Indeed if the PW Thesis is false, some epistemically possible statements will not be verified by any centered world, so the Core Thesis will be false.

It follows that the Core Thesis is true on the epistemic view of scenarios, and is true on the world-based view iff the PW thesis is true. I think there is good reason to hold that the PW thesis is true; but even if it is not, we may simply adopt the epistemic view of scenarios. Either way, the epistemic understanding of two-dimensional semantics plausibly yields an understanding of 1-intensions that satisfies the Core Thesis.

Given that epistemic intensions satisfy the Core Thesis, there are a number of natural applications that I will simply summarize here.

(i) Fregean sense: Because they satisfy the Core Thesis, epistemic intensions also satisfy the Neo-Fregean Thesis: ' A ' and ' B ' have the same intension iff ' $A == B$ ' is a priori. So epistemic intensions behave broadly like a sort of Fregean sense, tied to the rational notion of apriority. There are some differences. First, sentence-level Fregean senses are supposed to be true or false absolutely, but sentence-level epistemic intensions are true or false relative to a speaker and time (witness 'I am hungry now'). Second, apriority is weaker than cognitive insignificance, so epistemic intensions are less fine-grained than Fregean senses.[*] Nonetheless, it is clear that epistemic intensions serve as a broadly Fregean semantic value. (See Chalmers 2002s.)

*[[One can adapt the current framework to yield a more fine-grained sort of epistemic intension, by starting with a less idealized notion of epistemic possibility and proceeding from there. See Chalmers 2002e for more on this.]]

(ii) **Narrow content:** One can extend the current framework from language to thought in an obvious way. One can define epistemic intensions for beliefs and thoughts in the manner suggested in 3.8. The result can be seen as a sort of content of thought. It is very plausible that what results is a sort of *narrow* content, such that two physical and phenomenal duplicates will have thoughts with the same epistemic intension. (This narrowness is grounded in the narrowness of deep epistemic possibility: if a thought is a priori, then the corresponding thought of a physical and phenomenal duplicate will also be a priori.) This sort of content much more closely tied to cognition and reasoning than "wide content", and is well-suited to play a central role in explaining behavior. (See Chalmers 2002c.)

(iii) **Modes of presentation:** In analyzing the behavior of belief ascription, it is common to appeal to a notion of "mode of presentation", but there is little agreement on what sort of thing a mode of presentation is. Schiffer (1990) suggests that a mode of presentation must satisfy "Frege's constraint": roughly, that one cannot rationally believe and disbelieve something under the same mode of presentation. Because they satisfy the neo-Fregean thesis, epistemic intensions satisfy Frege's constraint perfectly, at least if one invokes an idealized notion of rationality that builds in arbitrary a priori reasoning. So it is natural to suggest that modes of presentation are epistemic intensions. In this way, one can use epistemic intensions to analyze ascriptions of belief.[*] (See Chalmers 2002c.)

*[[The current framework is compatible with a number of different proposals that give modes of presentation a role in belief ascription. I think it is particularly well-suited to so-called "hidden-indexical" accounts (Schiffer 1990). At a first approximation, it is plausible that 'X believes that S' will be true if the subject specified has a belief whose subjunctive intension is that of S, and whose epistemic intension falls into a certain "S-appropriate" class, which may be contextually specified. See Chalmers 2002c.]]

(iv) **Indicative conditionals:** One can use epistemic intensions to give a semantics for indicative conditionals that parallels in certain respects the common possible-worlds semantics for subjunctive conditionals. As a first approximation, one can suggest that an indicative conditional 'If S, then T' uttered by a subject is correct if the epistemically closest scenario that verifies S also verifies T, where epistemic closeness will be defined in terms of the beliefs or knowledge of the subject. (See Chalmers 2002t; see also Weatherson 2001 for an independent development of a similar idea.)

(v) **Conceivability and possibility:** The Core Thesis makes possible a certain sort of move from conceivability to possibility. If we say that S is conceivable when its negation is not a priori, then when S is conceivable, there will be a scenario verifying S. If we understand scenarios as possible worlds and the PW thesis is true, then when S is conceivable, there will be a centered possible world verifying S. This makes it possible to move from epistemic premises to modal conclusions, as is often done. Of course it is possible to embrace the current framework while rejecting the PW thesis and so rejecting the relevant move from conceivability to possibility. But the current framework at least shows how a certain sort of link between conceivability and possibility is tenable in light of the Kripkean phenomena that are often thought to be the greatest threat to such a connection. (See Chalmers 2002p.)

4 Epistemic intensions and contextual intensions

We have seen that there are two very different ways of understanding two-dimensional semantics: the epistemic understanding and the contextual understanding. On the epistemic understanding, 1-intensions are constitutively tied to the epistemic domain and satisfy the Core Thesis. On the contextual understanding, 1-intensions are not constitutively tied to the epistemic domain and do not satisfy the Core Thesis.

It is useful to examine the relationship between the two in somewhat more depth. First I will examine how the epistemic understanding deals with the problems that arise for the contextual understanding. Then I will examine to what grounds the resemblance of certain contextual intensions to epistemic intensions.

4.1 Problem cases

The first main problem area for contextual intensions involved sentences such as 'Language exists', which are a posteriori, but have a necessary contextual intension. These problems arose because contextual intensions require a token of the evaluated expression in the evaluated world. There is no such requirement for epistemic intensions, so the problem does not arise.

For example, there will be many language-free scenarios: there are many centered worlds with no languages, and there are many epistemically possible hypotheses under which there are no languages. If *D* is a canonical description of such a scenario, *D* will verify 'language does not exist'. Intuitively, if we consider such a scenario *W* as actual, we can say that *if W is actual, then language does not exist*. So the epistemic intension of 'language exists' will be contingent, as required.

The same goes for 'words exist', and something similar applies to 'I am uttering now'. In the latter case, there will be many centered worlds in which the subject at the center is not uttering, and there will be many epistemically possible hypotheses (for me) under which I am not uttering. If *D* is a canonical description of such a scenario, *D* will verify 'I am not uttering now'. So this expression will also have a contingent epistemic intension. The applies even to 'I am thinking now'.

'I exist' is a slightly trickier case. If 'I exist' is a priori, there is no problem. If 'I exist' is a posteriori (as I think is the case), then there will be various epistemically possible hypotheses for me under which I do not exist: for example, a hypothesis under which nothing exists (which is arguably itself not ruled out a priori). So on the epistemic view, there will be corresponding scenarios that verify 'I do not exist', and 'I exist' will have a contingent epistemic intension, as required.

On the world-based view, there is a worry: one might think that any centered world will verify 'I exist', since there is always a subject at the center. This raises a subtlety. In the general case, centering is *optional*: on the world-based view, the space of scenarios contains worlds without a marked subject and time, and perhaps worlds with only a marked subject or only a marked time. A world without a marked subject will then verify 'I do not exist'. The exact choices here will depend on exactly which indexical claims one holds to be a priori, but it should be possible to arrange things so that there is a verifying

centered world for every epistemically possible claim.

In any case, we see that problems that arise due to the required presence of a token do not arise here. At most there are problems due to the required presence of a *subject* in a centered world; but these will not arise on the epistemic view, and can be dealt with reasonably straightforwardly on the world-based view. So the epistemic understanding does not suffer from the problems of the contextual understanding here.

Another problem, at least for orthographic contextual intensions, concerned worlds where the subject at the center uses 'water' to mean something different, such as steel, so that the 1-intension picks out steel there, which is not the desired result. Again, this problem will not arise for epistemic intensions. In general, to evaluate the epistemic intension of 'water' at a scenario, the presence or absence of tokens of 'water' in that scenario will be irrelevant (with one qualification to be outlined shortly). What the epistemic intension of 'water' picks out in a Steel Earth scenario will depend on a number of other factors, especially the appearance and behavior of substances located around the center of the scenario, but there is no danger that it will pick out steel.

Note that this analysis requires that "My term 'water' means water" and similar claims are not a priori. If such a claim was a priori, then because a canonical description of the Steel Earth scenario will contain something like "My term 'water' means steel", the scenario would verify 'water is steel', which is the wrong result. But it is independently plausible that these claims are not a priori - at least if "'water'" is understood in purely orthographic terms. It is a posteriori that the string has any meaning at all, and it is a posteriori that it means what it does. If "'water'" is understood in partly semantic terms, so that it is constitutively tied to a given meaning for 'water', then the claim in question may be a priori; but this is not a problem, since in this sense the Steel Earth scenario will not verify "My term 'water' refers to steel". For more on this matter, see Chalmers (2002p) and Yablo (2002).

Note also that there is one case where evaluating an expression's epistemic intension may turn on the presence of tokens of that expression in a world: expressions used *deferentially*. It may be that Leverrier uses 'Neptune' to (rigidly) pick out whatever her husband refers to as 'Neptune'. If so, then in a given scenario, the epistemic intension of 'Neptune' will pick out roughly the referent of her husband's term 'Neptune' in that scenario (abstracting away from issues about the epistemic intension of 'my husband', etc). In this case, something like "If Neptune exists, my husband refers to it as 'Neptune'" will be a priori for her. Something similar to this will apply to other terms used deferentially, such as a non-expert's use of 'arthritis', although the details may be less clean. But here, we get only the results that would be expected. For example, if I use 'water' wholly deferentially, then if I consider as actual a Steel Earth scenario where those around me use 'water' for steel, then this scenario verifies 'water is steel' for me. This seems correct: for a differential user, although perhaps not for a nondeferential user, 'water is steel' expresses an epistemically possible thought.

(Note that even in deferential cases, evaluation turns on the referent of *others'* use of the expression. It may be that evaluation could also turn on one's own past use of the expression; but it cannot happen that evaluation will turn on the referent of one's own current use of the expression, since such a circular criterion cannot secure a referent. (I set aside pathological nonreferring cases, such as 'the referent of this

expression'.) So even in a strongly deferential case, the epistemic intension of 'water' will not turn on the referent of a use of 'water' at the very center of a scenario.)

There is also the Twin Earth case, where Twin Oscar uses 'water' to refer to XYZ. This was a problem for linguistic and semantic contextual intensions, since these are arguably not defined at such a world, whereas we would like the 1-intension of Oscar's term to return XYZ at this world. Again, the epistemic framework handles this unproblematically. The epistemic intension of 'water' returns XYZ at this world, *not* because Twin Oscar's term 'water' refers to XYZ (Twin Oscar's term is irrelevant), but because the scenario verifies the claim that XYZ has a certain appearance, behavior, relation to oneself, and so on, which in turn verifies 'water is XYZ'.

Finally, there was the problem of Fregean typing. It seemed that in order for contextual intensions to give roughly Fregean results, then one had to classify expression tokens under some sort of Fregean type. For a semantic contextual intension to give the right results, for example, one needed to appeal to some sort of prior Fregean semantic notion, which is unhelpful in the current context. No such problem applies to epistemic intensions. Because these intensions do not rely on tokens of the same type being present within scenarios, there is no need to isolate the common type under which these tokens fall. All one needs is the expression token itself, and its epistemic properties. This approach may *ground* an account of a sort of Fregean semantic value, but it need not presuppose any such account.

These advantages of the epistemic account over the contextual account are all grounded in the fact that the contextual understanding is an essentially metalinguistic understanding, while the epistemic understanding is not. The contextual understanding concerns content that an expression might have had; but the epistemic understanding reveals aspects of the content that it has. Everything is grounded in certain first-order epistemic claims, which we use as tools to reveal an expression's content, just as in the familiar modal case, various first-order subjunctive claims are used as tools to reveal an expression's content. As before, the cases are parallel.

4.2 Semantic contextual intensions

We saw earlier that some versions of a semantic contextual intension presupposed a quasi-Fregean notion of content. We can now turn the picture the other way around, and using the quasi-Fregean notion of content developed here to ground a semantic contextual intension.

Let us say that an *epistemic contextual intension* of an expression is the semantic contextual intension that derives from the use of epistemic intensions as the relevant semantic value. The epistemic intension of an expression token is a function from centered worlds to extensions, defined at worlds that have a token at the center with the same epistemic intension as the original token, and returning the extension of that token in that world.

It is easy to see that at the worlds where it is defined, an expression's epistemic contextual intension yields the same extension as its epistemic intension. If W is a centered world containing a token S' with

the same epistemic intension as the original token S: let E be the extension of S'. Then the epistemic contextual intension of S returns E at W. Further, the epistemic intension of S' (on the world-based view of scenarios) returns E at W, since W is actualized at S'. By identity of epistemic intensions, the epistemic intension of S also returns E at W. So S's epistemic contextual intension and epistemic intension are coextensive at W. Something similar applies on the epistemic view of scenarios, if we invoke the scenario corresponding to the centered world W.

So an expression's epistemic contextual intension is a restriction of the term's epistemic intension. For this reason, it will give appropriate quasi-Fregean results in many cases. It will not satisfy the Core Thesis: it will have the usual problems with 'language exists' and other metalinguistic claims, as it will not be defined at scenarios without the token at the center, or where the token has a different content. But it will be reasonably close for many purposes.

One could also define a epistemic version of the *cognitive* contextual intension of a token, defined at all worlds centered on a *concept* or *thought* with the same epistemic intension as the token, and returning the extension of that concept or thought. This would again be a restriction of the expression's epistemic intension, but it would be less of a restriction, since it would not require a linguistic token in the evaluated world. The Core Thesis will still be false due to various metacognitive claims and the like; but it will not be far off. I think this last notion is the best approximation that a contextual intension can yield to a quasi-Fregean content that satisfies the Core Thesis. It is clear, however, that this notion is essentially derivative of that of an epistemic intension.

This way of seeing things also helps to explain why some other contextual intensions give approximately Fregean results. It may be that there are various other features of type F of a subject or a token, that at least correlate with an epistemic intension to some degree. We can then set up an F-based contextual intension, defined at worlds centered on a subject or token with the same F features as the original, and returning the extension of the relevant token. Then in each such world, the token at the center will have at least approximately the same epistemic intension as the original token, and so in most cases will return the same or similar extension at that world. So the F-based contextual intension will approximate the behavior of a restriction of the original epistemic intension.

This applies especially to some cognitive contextual intensions. It may be that possession of a concept with a given epistemic intension is itself determined by features such as a concept's cognitive role and/or associated phenomenology, or more deeply by the subject's physical state, or functional state, or physical/phenomenal state. To know exactly which features are crucial would require a solution to the problem of intentionality, which is not yet available. But one can say: *insofar* as epistemic intensions are determined by features such as cognitive role or physical/phenomenal state, then corresponding contextual intensions (here, cognitive-role contextual intensions or physical/phenomenal contextual intensions) will be restrictions of the original epistemic intension, and so will behave in a quasi-Fregean manner. Again, however, the epistemic intension is the more fundamental notion of content.

4.3 Linguistic contextual intensions

We saw earlier (in section 2.2) that for some expressions, a linguistic contextual intension behaves in a quasi-Fregean manner. We are now in a position to see why this is.

For some expressions, their epistemic intension is part of (or determined by) their linguistic meaning. That is, some linguistic expression types are such that every token of that type has the same epistemic intension. As noted in section 3.7, something like this appears to apply to some pure indexicals, such as 'I', 'now', and 'today', to some descriptive terms, such as 'circular', and to some descriptive names, such as 'Jack the Ripper'.

When an expression token's epistemic intension is part of its linguistic meaning, then the token's linguistic contextual intension will be a restriction of its epistemic intension. This can be seen by the same sort of reason as in the previous section. Or one can simply apply the point there directly: if any token of S's linguistic type has the same epistemic intension, then S's linguistic contextual intension will be a restriction of its epistemic contextual intension, which is a restriction of its epistemic intension.

It follows that in cases such as 'I', 'now', 'circular', and 'Jack the Ripper', the terms' linguistic contextual intensions will be quasi-Fregean. They will not satisfy the Core Thesis because of the restriction to worlds containing relevant tokens, but they will be reasonably close. This explains the phenomenon noted in section 2.2: the quasi-Fregean behavior is a direct consequence of the fact that for these tokens, epistemic intension is an aspect of linguistic meaning. Once again, a contextual intension is interesting largely because of the degree to which it approximates an epistemic intension.

5 Other two-dimensional ideas

With this analysis of the contextual and epistemic understandings on the table, we are now in a position to turn to existing two-dimensional proposals to see how they fit into this analysis, and to use this analysis to help understand their foundations. I should say at the start that although I will occasionally criticize these approaches and argue that the approach I have recommended has certain advantages, any advantages are due largely to building on the insights that these approaches embody.

5.1 Stalnaker's diagonal

The diagonal proposition of Stalnaker (1978) is characterized as follows. We start with an understanding of propositions as sets of possible worlds, and with the idea that any utterance has a proposition as its content. (This propositional content coincides roughly with what I have called a subjunctive intension.) We can then say: the utterance could have had different propositional content. So there are worlds where the utterance has different propositional content. This allows us to define an utterance's *propositional concept*, which is a function from possible worlds to propositions, defined at any world containing the utterance, returning the propositional content of the utterance at a world. We can then define the utterance's *diagonal proposition* as the set of worlds such that the utterance's propositional concept, evaluated at that world, yields a proposition that is true at that world.

As defined here, a diagonal proposition is much like a token-reflexive contextual intension. There are minor differences. A token-reflexive contextual intension was defined directly in terms of what an utterance's truth-value would be at a world, rather than in terms whether the proposition it expresses would be true at that world, but it is clear that within the propositional framework, these yield the same results. A diagonal proposition is a set of possible worlds or equivalently a function from worlds to truth-values, whereas a token-reflexive proposition was a function from centered worlds to truth-values. But again, that is a minor difference: one can translate between the relevant worlds and centered worlds either by "marking" the location of the token as a center, or by removing the marked center from the location of the token. (Token-reflexive contextual intensions uniquely do not need a center to specify the relevant token, since the token is independently identified by a sort of transworld identity with the original token.) So diagonal propositions are equivalent to token-reflexive contextual intensions.

The behavior of a token-reflexive contextual intension is not clear unless we know which properties are essential to a token and which properties are inessential. It is clear that Stalnaker holds that a token's semantic properties are not essential to it, since he holds that it could have had different semantic content. It seems plausible that on his picture, a token's orthographic properties are essential to it; at least, in all examples, a token's orthographic properties are held constant across its possible occurrences, so I will assume this in what follows. It is not entirely clear which other properties are essential (language? probably not. speaker? maybe), but this is enough to be going on with.

>From what we have said here, it appears that a token's token-reflexive contextual intension will be a restriction of its orthographic contextual intension, restricted to cases where the orthographically identical token is the original token. One might think that it's general behavior will be very much like that of an orthographic contextual intension: for my utterance 'water is H₂O', there will be worlds at which it means that steel is orange; if so, its token-reflexive contextual intension will be defined and presumably false there.

However, this sort understanding is inconsistent with a central point in Stalnaker's original paper, where he says that for a sentence to have a necessary diagonal proposition is for it to be an a priori truth. He also says that an official's utterance of 'this bar is one meter long' could not have expressed a false proposition. It is unclear how this can be justified. Why could not the utterance have meant something like "that boat is two miles long", and been false? It is natural to suppose that that Stalnaker is holding fixed some intuitive sort of meaning and content. (Thus yielding something that behaves like a semantic/orthographic contextual intension, for an intuitive Fregean semantic value.) But it is not clear what aspect this could be: the only content that his account officially recognizes is propositional content, which is explicitly held to vary with possible occurrences of an utterance; and even if there were some further aspect of content, it is unclear why this sort of content should be essential to an utterance and the other sort inessential. Alternatively, it may be that Stalnaker is assuming that some sort of cognitive factor (e.g. associated cognitive role?) is essential to an utterance (thus yielding something that behaves like a cognitive/orthographic contextual intension), but this is nowhere specified. So the claim about apriority and about the meter case appears to be ungrounded within the article.

In later work, Stalnaker retreats from the claim about apriority (although to my knowledge he does not

address the discussion in the original article), and allows a much wider range of behavior for an utterance across possible worlds. For example, in Stalnaker (1999) he allows where that there are worlds where 'Julius', used in the actual world as a descriptive name for the inventor of the zip, is used instead for the inventor of bifocals. And in Stalnaker (2001) he allows that there are worlds where our word 'tiger' refers to pieces of furniture. In both of these articles he explicitly denies that necessity of diagonal proposition corresponds to apriority, as seems reasonable. In effect, this diagonal proposition behaves very much like an orthographic contextual intension, or a straightforward restriction thereof.

As such, a diagonal proposition is clearly useful. For example, Stalnaker often uses diagonal propositions to model situations of communication, in which a hearer hears an utterance but is unsure what the speaker meant by it, or has false beliefs about what the speaker meant by it. It seems clear that this sort of metalinguistic use requires something quite different from a quasi-Fregean notion, so this is reasonable. It is less clear that diagonal propositions are useful for addressing matters of cognitive significance, rational inference, apriority, and the like, especially when divorced from issues about communication. Stalnaker sometimes suggests this sort of use, but I think the grounds here are weaker.

For example, in a recent paper (2001), Stalnaker holds that the "metasemantic" framework with diagonal propositions can "provide an explanation for the phenomena that Kripke's work brought to light" - where this phenomenon is the distinctive behavior of the class of a posteriori necessities such as 'Hesperus is Phosphorus', 'water is H₂O', and so on. This is a surprising claim. What is most distinctive about these phenomena are the *differences* with standard necessities such as 'all bachelors are unmarried', '2+2=4', and so on. If diagonal propositions function as Stalnaker (1978) suggests, the distinction would be straightforwardly represented by the fact that the second class have necessary diagonal propositions and the first do not. But on the metasemantic understanding, there seems to be no way to draw the distinction using diagonal propositions alone. For *both* sorts of necessity, there will be many worlds at which the diagonal is false, and there are no clear patterns that are distinctive to the first class. So it is not clear how the "explanation" is supposed to work.

In a brief ensuing discussion of 'Hesperus is Phosphorus', Stalnaker appeals to the fact that there is a world where the diagonal proposition is false. But clearly this holds equally for 'all bachelors are unmarried'. One might find some differences if one focuses on a *restriction* of the diagonal proposition. It is notable that in the worlds Stalnaker discusses, 'Hesperus' and the like appear to be used with the same reference-fixing intentions as the original term, for example. It may be that under this sort of restriction, the two sorts of necessities behave differently: there are counterexamples to the diagonal for one class but not the other. But the restriction is doing all the work: in effect, it invokes something more like a cognitive contextual intension, rather than a diagonal proposition per se. Such intensions may be able to model the Kripkean distinction at least approximately, if imperfectly. Stalnaker appears to use similar tacit restrictions in some other cases; in all these cases, it seems that in effect a restricted contextual intension does the explanatory work.

The explanatory power of restricted contextual intensions here itself plausibly derives from that of epistemic intensions. Epistemic intensions handle these phenomena straightforwardly: for Kripkean necessities, there is a falsifying scenario, and for standard necessities there is not. For reasons we have

seen, certain restricted contextual intensions approximate epistemic intensions, and so approximate this behavior (with some exceptions). So it is plausible that the usefulness of diagonal propositions in this context derives indirectly from the usefulness of epistemic intensions.

Stalnaker contrasts his "metasemantic" version of the framework with "semantic" versions, on which 1-intensions are an aspect of semantic content, and suggest that the apparent attractions of the latter in explaining these phenomena derive from the attractions of the former. The above suggests that this is not quite right: the attractions of Stalnaker's version of the framework in this domain derive from the attractions of the epistemic version. As for whether the epistemic understanding is itself a "semantic" understanding: this matter depends on what is meant by "semantic". Stalnaker mostly uses the term to contrast with "metasemantic", indicating an aspect of first-order content rather than a metalinguistic dependence: in this sense, epistemic intensions are semantic. Stalnaker also sometimes uses the term to indicate those aspects of content that are built into linguistic expression types, rather than varying across tokens: in this sense, epistemic intensions are not semantic. Stalnaker appears to assume that his opponent's framework is semantic in both these senses;[*] but these are very different distinctions. Epistemic intensions need not be built into linguistic meaning to be a sort of first-order content that does explanatory work.

*[[The only opponent that Stalnaker cites is Chalmers (1996). I note that Chalmers (p. 58) explicitly leaves open the question of whether different speakers might associate different 1-intensions with the same word. Stalnaker also argues that the framework cannot yield an account of the a priori; I agree, and have not claimed that it can.]]

In any case, I think it is clear that diagonal propositions and epistemic intensions both have useful roles to play. Diagonal propositions are best suited to analyzing matters of context-dependence, and epistemic intensions are best suited to analyzing the epistemic domain.

5.2 Kaplan's character[*]

*[[I should note that I am less confident about my discussion of Kaplan than about any other section of this paper. Suggestions are welcome!]]

Kaplan's notion of character is set out as follows. We assume a prior notion of the proposition expressed by an utterance: such a proposition is something in the vicinity of a 2-intension, although it may be a singular proposition instead. For some linguistic expression types (e.g. 'I'), utterances of the same type can express different propositions in different contexts. The character of an expression type is a function from contexts to propositions, returning the proposition that an utterance would express in a given context.

At first glance, it may seem that character is much like a linguistic contextual intension. There are some superficial differences. For example, Kaplan's contexts are not quite centered worlds; but in general they include a few parameters (speaker, time, etc) as well as an 'actual-world', and so they can be modeled by centered worlds. Also, character is a function from contexts to propositions, not to extensions. But one

can diagonalize character by evaluating the proposition associated with a given context in the world of that context, yielding an associated function from contexts to extensions.

In specific cases, character behaves quite like a linguistic contextual intension. We have seen that the linguistic contextual intension of indexicals such as 'I', 'now', and 'today' pick out the speaker, time, day (and so on) of the center of all worlds at which they are defined. The same is true for (diagonalized) character, on Kaplan's account). At the same time, we have seen that the linguistic contextual intension of a name arguably picks out the same individual at all worlds where it is defined. Again, the applies to character, on Kaplan's account.

One case where the two apparently behave differently is for demonstratives such as 'that'.[*] If I use 'that' intending to refer to an object in front of me, then its character will pick out (roughly) an object in front of the speaker in all contexts. But the linguistic contextual intension will not: what it picks out in a context will depend on the underlying demonstration or intention of a speaker in that context. But in a way, this is the exception that proves the rule. In Kaplan's formal analysis, he stipulates that different uses of 'that' (roughly, those corresponding to different demonstrations) are tokens of *different* words: 'that1', 'that2', and so on. Under this stipulation, it is plausible that a linguistic contextual intension for one of these instances of 'that' will behave as characterized above.

*[[I will use "demonstrative" for expressions such as 'that', 'he', and 'you', while using "indexical" for expressions such as 'I', 'here', and 'now'.]]

However, there are aspects of Kaplan's discussion that make it clear that character is fundamentally different from a linguistic contextual intension. Kaplan stresses that when we evaluate a sentence's character in a context, we do not evaluate an *utterance* of that sentence within the context. Rather, we evaluate an *occurrence* of the sentence at the context. An occurrence is in effect an ordered pair of a sentence and a context. And crucially, the context need not itself contain an utterance of the sentence. In effect, this is to allow that the character of an expression can be evaluated directly at a centered world, whether or not the world contains a token of the original expression or not.

Kaplan's reason for doing this are largely tied to his desire for a *logic* of demonstratives. He suggested that arguments involving demonstratives should be *valid* in virtue of their character: that is, a conclusion should follow from premises in virtue of an appropriate relation among their characters. But if the character of a claim were restricted to contexts containing an utterance of that claim, then each premise and the conclusion would be defined across different contexts, so their characters could not stand in the right sort of relation. He also says that there are sentences that express a truth in certain contexts, but in no contexts in which they are uttered: e.g. 'I say nothing'. If so, contexts cannot be required to contain a token of the relevant utterance.

For these purposes, it is natural to suggest that character should be something more like an epistemic intension. Validity, at least as Kaplan uses it here, is a deeply epistemic notion, tied to apriority and to rationally compelling inferences (in Kaplan's discussion, it is clear that validity is not constitutively tied to necessity). The sort of intension that is *constitutively* tied to validity and to apriority is an epistemic

intension. Similarly, for the intension of 'I say nothing' to be false in the relevant contexts, the best candidate is a something like an epistemic intension.

It is difficult to adjudicate what Kaplan intends, however, since he never specifies *how* to evaluate an expression's character in a context. He simply stipulates that expressions have a character associated with them, and then discusses character's properties. He does say on some occasions that character picks out what the expression would pick out if uttered in that context, but he retracts this because of the point about occurrences (though presumably *when* a context contains a relevant utterance, character still returns what the utterance picks out). He also says (p. 505) that character is set by linguistic conventions and determines the content in a context, and he suggests that character is determined by a demonstration (pp. 526-27) or a directing intention (pp. 587-8). But nothing here tells us how to evaluate character in contexts not containing the utterance. In some cases the matter seems reasonably straightforward: 'I' picks out the marked subject in a context, 'you' picks out a marked addressee. But there seems to be no general principle here for assigning an evaluation function to an expression type.[*]

*[[Another complication is that is not entirely clear what is built into the relevant context. On a couple of occasions (p. 528, p. 588) Kaplan entertains the idea that a context explicitly contains a parameter, for a demonstratum to serve as referent. If this is done, it renders the question about how to evaluate character trivial, at cost of trivializing many other aspects of the framework. It also removes any special role for demonstrations and directing intentions in contextual evaluation, and removes the deep connection with cognitive significance. Partly for these reasons, and partly because it eliminates the connection between demonstratives and indexicals, this seems not to be Kaplan's considered view. So presumably a directing intention does the work of picking out a referent in a context in which that referent is not explicitly specified.]]

One way to address the question is to ask: is it *constitutive* of character that validity and apriority are governed by character? Or is this merely a feature that character turns out to have? It seems that this cannot be constitutive, for the obvious reason that in the case of proper names, validity and apriority come apart from character. But now the question arises: what justifies Kaplan's claim that the character of indexicals and demonstratives must be logically well-behaved? In effect, it is this that determines his treatment of occurrences. If the notion of character were constitutively connected to validity, this would be reasonable. But since it clearly is not, this seems less reasonable. If character is independently grounded, it seems that one might equally say: character is reasonably close to reflecting validity and the like, but unfortunately it is imperfect.

Perhaps the most likely diagnosis is: the starting notion of character is not constitutively connected to the domain of validity and apriority. (Perhaps it is something like a linguistic contextual intension.) But at least in the case of indexicals and names, this notion of character turns out to come very close to reflecting this domain. It turns out that a slight modification makes the correspondence precise, so at least in this case we adopt the modified notion. The resulting notion appears to be something quite close to an epistemic intension. It is not exactly an epistemic intension, for example because of the use of further parameters in a context. But it seems to behave in a quite similar way.

This raises the question: why not do the same for names? If character is to be connected to apriority, why

not make character behave something like an epistemic intension in this case? The initial answer is that Kaplan thinks that names do not behave this way: their contents are essential to them, so they do not pick out different contents in different contexts. In discussing the matter, Kaplan notes especially (p. 562) that occurrences of 'Aristotle' that refer to different people are different words. One might respond that this would be relevant if we were defining the *contextual* intension of names, but we are now dealing with a modified notion. The fact that a name has its referent essentially (and the point about 'Aristotle') is compatible with its *epistemic* intension picking out different referents in different contexts.[*] But the crucial point may in fact be something different: character is supposed to be a sort of linguistic meaning, but the names as linguistic types do not have epistemic intensions (at best, epistemic intensions vary between tokens).[*] So character cannot be epistemic intension.

*[[Something I have not mentioned is Kaplan's 'Fregean theory of demonstrations', arguing that a demonstration does not have its referent essentially. It seems that this point would be highly relevant to a contextual understanding of demonstratives, but it is not so clear that it is required for an epistemic understanding.]]

*[[This sort of point about the difference in cognitive significance between different tokens of a name is never mentioned explicitly in Kaplan's article, but it may be playing a role implicitly in his claim that names do not have nontrivial character semantically associated with them. A useful diagnostic question would be whether descriptive names (if there are any), such as 'Jack the Ripper', can have nontrivial character. If yes, then variability is plausibly the key reason standard names have trivial character. If no, then essentialness of referent is plausibly the key reason.]]

Still, an obvious response is that the same holds for demonstratives. Kaplan's formal move of stipulating that different tokens of a demonstratives are different words is clearly something of a convenient trick: the force is to suggest that character need not really be associated with a linguistic type, but with a token.[*] If so, then we could say the same for names, perhaps making a similar stipulation, or perhaps not. It is not clear exactly how the cases are relevantly different. One suggestion is that different tokens of a name seem to be more closely tied than different tokens of a demonstrative, with some sort of associated assumption of communication, agreement and disagreement, and so on. If so, then assigning all these name tokens a different linguistic type might be even more counterintuitive than for demonstratives. But it is not clear exactly what the rules are here. And it is not clear why one should not take the real moral of the demonstrative case to be that character is fundamentally a property of tokens rather than linguistic types, in which case it is no longer obvious that names must have trivial character.

*[[See Braun 1996 on this topic. Braun notes that Kaplan also seems to take the informal strategy of taking 'that' to be a single word, and of associating character not with the word but with a word-plus-demonstration pair. This raises the question: since the relevant demonstrations (especially according to the later Kaplan) are a sort of intention, why not analogous associate character with a name-plus-intention pair? Then one will in effect have character for utterances of names.

One might even note: for a use of a name, there can be a directly linked demonstrative. This can happen with "John...he...", or better, simply with "he" while this is simply backed by the intention to refer to John (perhaps this can be a mild counterfactual variant on the original utterance). It does not seem entirely unreasonable to say that this last demonstrative has nontrivial character. If so, one could use this character to motivate a nontrivial character in the vicinity of any token of 'John'. If not, then it seems to follow that the Fregean theory of demonstrations is false in at least some cases, and one wants to know more about the rules for associating character with demonstratives and demonstrations.]]

In any case, my best guess as to what constitutes character is the following: character is something like an epistemic intension in cases where it is reasonable (perhaps at a stretch) to assign epistemic intension to a linguistic type. If not, because of variability of epistemic properties between tokens, then character is something else, perhaps more like a contextual intension, or perhaps stipulated to be a function that returns an expression's actual content (whatever it is) at all contexts. (If character really is something like "epistemic intension insofar as it is associated with a linguistic type", one might equally simply say that names have no character, rather than saying that they have constant character. That does not seem wholly contrary to the spirit of Kaplan's discussion.)

(There are a couple of interesting diagnostic cases. First, what is the character of a descriptive name, such as 'Jack the Ripper'? Second, what is the character of a context-sensitive predicate such as 'heavy'? Kaplan is silent about these matters, but the behavior of character here might help decide among the alternatives above.)

The distinction between contextual and epistemic understanding also helps bears on a recent controversy about occurrences. Garcia-Carpintero (1998) argues for the superiority of a Reichenbachian token-reflexive account of indexicals over an account that relies on Kaplan's occurrences, in effect suggesting that a sort of token-reflexive contextual intension (requiring the token in a context) is truer to the data than an account that does not require tokens. As part of his argument, he denies that there is any reasonable intuition that there are contexts in which 'I am not uttering now' is true. Our discussion makes it possible to render a split verdict here. On a contextual understanding of evaluation in contexts, there is on such intuition. But on an epistemic understanding, there is such an intuition. The intuition, I think, is that 'I am not uttering now' is not false a priori, so that there are epistemic possibilities in which it is true. These epistemic possibilities are scenarios in which the subject at the center is not uttering. The difference between Kaplan's and Reichenbach's frameworks may then be grounded in the fact that Kaplan's semantic value for an indexical is constitutively tied to its epistemic properties, while Reichenbach's is tied to its contextual properties.[*]

*[[Note, though, that on the framework I have suggested, epistemic intensions are fundamentally assigned to utterances, not to occurrences. Occurrences play a role in that epistemic intensions are *evaluated* at scenarios that need not contain the relevant utterance. This suggests that there are two issues dividing the Reichenbachian and the Kaplanian, involving the assignment and the evaluation of the relevant semantic values, and that these two issues should be separated.]]

In any case, it seems plausible that there are elements of both the contextual understanding and the epistemic understanding in Kaplan's account, not always disentangled. Perhaps character is fundamentally an extension of a linguistic contextual intension; perhaps it is fundamentally a sort of epistemic intension; or perhaps there is no fact of the matter. But it is clear in any case that much of the value of character in the case of demonstratives that in this case, it behaves much as an epistemic intension does (whereas in the case of names, it does not). It does not seem unreasonable to hold that character is useful for epistemic purposes precisely to the extent that it approximates or coincides with an epistemic intension.

5.3 Evans' Deep Necessity

In addressing Kripke's problems of the contingent *a priori*, Evans focuses on the case of descriptive names. He introduces the descriptive name 'Julius', whose references is fixed to whoever invented the zip. Then 'Julius invented the zip' seems to be *a priori*. In analyzing the case, Evans distinguishes between two sorts of necessity: "deep necessity" and "superficial necessity". Instances of the "contingent *a priori*" (such as 'Julius invented the zip') are superficially contingent but deeply necessary; instances of the "necessary *a posteriori*" are superficially necessary but deeply contingent. Evans says that whether a statement is deeply necessary or contingent depends on what makes it true; and whether it is superficially contingent depends on how it embeds under modal operators.

Superficial necessity is defined as follows. A sentence Q is superficially contingent if 'possibly $\sim Q$ ' is true, or equivalently, if there is some world W where Q is not true_W. Here, the possibility operator is clearly subjunctive possibility ("it might have been that"), and the possible-worlds evaluation is clearly subjunctive counterfactual evaluation of the Kripkean sort. So superficial necessity is a second-dimensional notion: S is superficially necessary when it has a necessary 2-intension or subjunctive intension.

Deep necessity and contingency are characterized in the following passage toward the end of Evans' article:

We have the idea of a state of affairs, or a set of state of affairs, determines by the content of a statement as rendering it true, so that one who understands the sentence and knows it to be true, thereby knows that such a verifying state of affairs exists. A deeply contingent statement is one for which there is no guarantee that there exists a verifying state of affairs. If a deeply contingent statement is true, there will exist some state of affairs of which we can say both that if it had not existed the sentence would not have been true, and that it might not have existed. The truth of the sentence will thus depend on some contingent feature of reality. (Evans 1979, p. 185)

This passage has a strong epistemic element in the first half; and a strong contextual element in the second half. To understand these we need to examine the discussion earlier in Evans' article.

Evans introduced the notion of the *content* of a sentence earlier as capturing an epistemic element. Evans says that when two sentences have the same content, they are *epistemically equivalent*: a person who understands both cannot believe what one says and disbelieve what the other says. Evans makes a strong distinction between the content of a sentence and the *proposition* expressed by a sentence, which is a function from possible worlds to truth-values of the sort associated with the modal contexts of superficial necessity. He notes that two sentences that express the same proposition can have different contents, and argues that two sentences with the same content can express different propositions: e.g. 'Julius is F' and 'The inventor of the zip is F'.

Evans holds that there is a notion of "making a sentence true" that is tied directly to content (which he

distinguishes from an alternative sense tied to proposition expressed). He says:

...if two sentences are epistemically equivalent, they are verified by exactly the same state of affairs, and what one believes in understanding the sentence and accepting it as true is precisely that some verifying state of affairs obtains. On this conception, the same set of states of affairs makes the sentence 'Julius is F' true as makes the sentence 'The inventor of the zip is F' true. If $x, y, z, \dots$ is a list of all objects, then any member of the set $\{x\text{'s being the inventor of the zip} \ \& \ x\text{'s being F}; y\text{'s being the inventor of the zip and } y\text{'s being F}; z\text{'s being the inventor of the zip and } z\text{'s being F} \dots\}$ will suffice to make the sentence true.
(Evans 1979, p. 180)

On this conception, making a sentence true, at least in the case of a descriptive name, seems to involve something like satisfying its epistemic intension. In the sense of 'verify' that I tied to epistemic evaluation, 'Julius invented the zip' will be verified precisely when the conditions that Evans suggests for "making the sentence true" obtains. The claim that epistemic equivalence entails verification (in Evans' sense) by the same states of affairs also suggests a tie to epistemic intension, and suggests a link between the two notions of 'verification'. If deep necessity is tied to 'making true' in this sense, then at least in the case of descriptive names it seems to be a sort of necessity of epistemic intension.[*]

*[[This is not to say that Evans' notion of content is in general a notion of something like an epistemic intension. There is nothing in Evans' account of content that suggests that a term's content is exhausted by its epistemic properties, and especially beyond the case of descriptive names, it may involve elements such as a referent essentially. It may also be that insofar as it applies to the case of names generally, Evans' notion of verification behaves quite differently from the notion I have defined. But at least in the case of descriptive names, there seems to be a close connection.]]

But Evans also characterizes this notion of 'making true' in alternative terms:

But there is an ineliminable modal element in the notion of what makes a sentence true. For what can it mean to say that any one of a set of state of affairs renders a sentence true, other than to say that, if any one of them obtains, the sentence will be true, and if any of them *had* obtained, the sentence *would have been* true.

On *this* characterization, a state of affairs makes a sentence true roughly when it satisfies a sort of *contextual* intension of the sentence. "Making true" is characterized in terms of a metalinguistic subjunctive about truth-values the sentence *could* have had. This strongly suggests something like a linguistic contextual intension, as sentences here correspond to linguistic types.

This understanding is reinforced by the claim that Evans goes to some length to argue that if y had invented the zip and had been F, y would have been the referent of 'Julius', and 'Julius is F' would have been true *as a sentence of English*. He argues that there is no *semantic* connection between 'Julius' and a particular referent, so one can suppose that the term could have had a different referent without supposing a semantical change in English. He says: "exactly the same theory of meaning serves to describe the language which would be spoken had y invented the zip, as describes the language which is actually

spoken" (Evans 1979, p. 182). If 'making true' is tied to a linguistic contextual intension, this (controversial) claim is crucial to arguing that making true behaves as Evans' suggests. If 'making true' were tied to something like an epistemic intension, this claim would be unnecessary. So this strongly suggests an understanding as a linguistic contextual intension.

How can all this be reconciled? My suspicion is that it cannot be. There are cases when the contextual understanding clearly gives the wrong results. If we let 'L' be a descriptive name for the number of actual languages: then on the first understanding described above (by analogy with what Evans says about 'Julius'), for a state of affairs to make 'L=0' true it suffices for the number of languages to be zero according to that state of affairs. But on the second understanding, to make 'L=0' true it will be required that the sentence 'L=0' be true according to that state of affairs; and clearly this cannot happen. So on the first understanding, there will be a verifying state of affairs, and 'L>0' will be deeply contingent; on the second, there will be no verifying state of affairs and 'L>0' will be deeply necessary.

One might also worry: Evans holds that the same states of affairs make true 'Julius is F' as 'The inventor of the zip is F', but it is not clear how the contextual understanding above can deliver this result, as the states of affairs invoked there seem to require the distinct sentences. One might suggest that Evans is tacitly appealing to counterfactuals such as 'If the state of affairs had obtained *and* the sentence had been uttered, it would have been true'. This would be in effect to invoke a sort of extended contextual intension, with the various unclarities that come along with that notion. We can also note that it does not help with the case of 'L' above.

The paragraph quoted at the start seems to contain elements of both understandings of deep contingency. One could argue that in this paragraph, the epistemic elements in the first half could be reconciled with an overall contextual understanding, since those elements explicitly make appeal to metalinguistic knowledge, in knowledge of a sentence's truth. But such an understanding is hard to reconcile with Evans' overall project, on which 'making true' is tied to content. Content is tied to something like apriority and epistemic equivalence, and these are tied to something like an epistemic intension, not a contextual intension.

So Evans account contains evidence of both an epistemic and contextual understanding; and there is a clear conflict between the two. My guess is that Evans' attachment to a sort of epistemic understanding of 'making true' and deep necessity for descriptive names runs deeper, and that the attempted clarification of the notion in contextual terms was a mistake that might be jettisoned, leaving a broadly epistemic picture intact. But it is hard to know for sure.

If one jettisons the contextual elements, the notions of 'making true' and of deep necessity that remain in Evans' discussion appear to be broadly compatible with an epistemic understanding, at least in the case of descriptive names. Of course Evans restricts his discussion to the case of descriptive names, and would probably resist an extension of his analysis to other names, as elsewhere there is on corresponding descriptive element in the language's semantics. It is hard to know what Evans would say about the status of 'Cicero is Tully': it is epistemically nontrivial, but it seems that no states of affairs could make it false, on Evans' semantic approach. It may be that Evans intended the notions of deep necessity and 'making

true' to be inapplicable here. But one could argue the seeds are nevertheless present for a more general account.

5.4 Davies and Humberstone's "Fixedly Actually"

Davies and Humberstone (1980) give a "formal rendering" of Evans' distinction between deep and superficial necessity, using independently motivated tools from modal logic developed in Crossley and Humberstone (1977). The formal framework starts with a necessity operator N , and supplements it with an "actually" operator A , meaning "it is actually the case that". This allows one to represent claims such as 'It is possible for everything which is in fact ϕ to be ψ ', as ' $\text{POSS all } (x) (A \phi(x) \rightarrow \psi(x))$ '. One naturally give a model theory for these notions by supplementing a the space of possible worlds needed for the necessity operator with a designated "actual world", where $A \alpha$ is true at a possible world iff α is true at the actual world.

This framework naturally suggests a further idea: just as one can ask whether α is true with respect to a possible world (holding the actual world fixed), one might ask whether α would be true if a *different* world were designated in the actual world. This notion is modeled by adding a further "fixedly" operator F , where $F \alpha$ is true at a world W iff α is true at W no matter which world is designated as actual. One can then derive the "fixedly actually" operator FA . Given the definitions of F and A , $FA \alpha$ will be true at a world if, when that world is designated as the actual world, α is true at that world.[*]

*[[It should be noted that Davies and Humberstone's official model theory for the system with F and A involves one-dimensional evaluation of the relevant propositions at a world, where there is an underlying space of worlds and designated actual world in this space. They note a model theory could also be given using the tools of two-dimensional (or doubly-indexed) modal logic, by holding that α is true at $(W1, W2)$ on the doubly-indexed version iff α is true at $W2$ on the singly-indexed version when $W1$ is designated as the actual world.

So it may not be strictly correct to say, as is often said, that the ideas of two-dimensional semantics are grounded in two-dimensional modal logic. Two-dimensional modal logic *per se* does not enter essentially into grounding the frameworks of Kaplan, Stalnaker, and Evans; and even in the case of Davies and Humberstone, two-dimensional modal logic is presented merely as an optional means of representation. Of course many of these ideas can be naturally represented using the tools of two-dimensional modal logic, and one there is plausibly a relationship between the conceptual bases of these frameworks and of two-dimensional modal logic.]]

Davies and Humberstone note that "fixed actually" can itself be seen as a sort of necessity, and suggest that it is closely linked to Evans' deep necessity. The formal framework can itself be used to represent some contingent a priori truths, such as ' $A \phi \leftrightarrow \phi$ ', where ϕ is a contingent truth. If we call this sentence P , then ' $FA P$ ' is true but ' $\text{necessarily } P$ ' is not. So FA seems to go along with the sentence's apriority in this case, and the behavior of the 'necessarily' operator seems to go along with its contingency. Davies and Humberstone suggests T is superficially contingent when ' $\text{necessarily } T$ ' is false, and deeply contingent when ' $FA T$ ' is false.

In this framework, the behavior of a term such as 'Julius' is directly reflected by that of a term such as 'the

actual F'. For example, 'if anyone invented the zip, Julius did' is closely akin to 'if anything has F, the actual F has F'. The both of these contingent a priori sentences have the same patterns of intuitive evaluation, and the second can be rendered in the formal language as holding fixedly actually but not necessarily. Davies and Humberstone suggests that all descriptive names such as 'Julius' are semantically equivalent to phrases such as 'the actual F', and can be modeled correspondingly by the formal system. In this way, the deep and superficial modal properties of sentences involving 'Julius' are straightforwardly explained.

Davies and Humberstone also suggest that this analysis might be applied to terms such as 'water', 'red', and 'good', at least on certain philosophical views. In all these cases, there is a "actually" involving descriptive expression in the vicinity that yields qualitatively similar behavior, at least with respect to matters of necessity and apriority; and the behavior of those descriptive expressions can be modeled using the formal framework described above. Davies and Humberstone stop short of asserting that the English terms are semantically equivalent to these descriptive expressions, but they note that there is at least a resemblance.

Davies and Humberstone are quite explicit that their account applies only to terms such that their reference is semantically fixed by a description. So it does not apply to proper (non-descriptive) names: in fact, they say that any true identity statement using proper names is both superficially and deeply necessary, since such names are not even implicitly 'A'-involving. They also note that if it became widely known that 'Julius' referred to a certain individual (Tom, say) and the term continued to be used, it would soon cease to be a descriptive name, and claims that were previously deeply contingent (e.g. 'Julius is Tom') would become deeply necessary. It is for this sort of reason that there are doubts about whether the framework applies to terms such as 'water', where there is acquaintance with the referent, and also knowledge of its underlying nature.

Does this framework involve a contextual understanding, an epistemic understanding, or neither? It is somewhat hard to say, since the logical apparatus is a formal framework that does not come interpreted in this sort of way. But in discussing specific examples, such as 'Julius is F', Davies and Humberstone hold that the FA version will be true in a world iff the inventor of the zip is F there, and no special role is given to a token of 'Julius'. In fact considerations about the truth of possible tokens is nowhere invoked in D&H's article, except in giving a truth theory for possible tokens of sentences *containing* F, A, and the like; but here the tokens simply play the role they do in any truth theory, and there is no suggestion that the truth of FA phi is constitutively tied to the truth of possible tokens of phi.[*] In this respect, it appears that they depart from Evans.[*]

*[[Rumfitt 19xx addresses the question of whether there is something metalinguistic about D&H's notion of "considering a world as actual", and also gives a negative answer. He notes that Davies and Humberstone's two's modal notions are supposed to be of the same semantic sort, and there is nothing metalinguistic about the standard notion of necessity, so there is plausibly nothing metalinguistic about the novel notion.]]

[[Davies and Humberstone (p. 16) do endorse Evans' remark that if a certain state of affairs had obtained, a sentence would have been true even though it is not true with respect to that situation. But here, they are

simply illustrating how possible tokens of sentences such as 'Grass is actually green' behave, and not giving a criterion for the modal evaluation of the sentence in a world.]]

It seems that the rules for evaluating a natural-language term in a world considered as actual" are something like the following: if the term is equivalent to 'the actual F', then pick out F; otherwise, pick out what the term would pick out at that world under standard modal evaluation. This suggests that in the case of descriptive names, evaluation here will work as it does on the epistemic understanding. One might hesitate to ascribe too strong an epistemic understanding here, given the treatment of proper names, and also because D&H do not at any point even tacitly invoke the idea of considering that world as an epistemic possibility. Still, at least in the case of a descriptive name, this framework and the epistemic understanding yield largely equivalent results.

5.5 Chalmers' primary intensions

Chalmers[*] (1996, pp. 56-65) presents what he calls a "synthesis of ideas suggested by Kripke, Putnam, Kaplan, Stalnaker, Lewis, Evans, and others".[*] He distinguishes what he calls the "primary intension" and the "secondary intension" of a concept. (A concept appears to be either a token expression or a mental item.) These are initially characterized as follows (p. 57):

*[[I will start by examining the text from the outside; hence the third-person reference. Autobiographical remarks are in footnotes and below.]]

*[[One moral of the discussion above, it is that a blanket citation of theorists who have worked on two-dimensional ideas has the potential to confuse more than it clarifies!]]

there are two quite distinct patterns of dependence of the referent of a concept on the state of the world. First, there is the dependence by which reference is fixed in the actual world, depending on how the world turns out; if it turns out one way, a concept will pick out one thing, but if it turns out another way, the concept will pick out something else. Second, there is the dependence by which reference in *counterfactual* worlds is determined, given that reference in the actual world is already fixed. Corresponding to each of these dependencies is an intension, which I will call the *primary* and *secondary* intensions, respectively.

The secondary intension seems to be the familiar sort of intension (2-intension, subjunctive intension) across possible worlds. The nature of the primary intension is somewhat less clear. The characterization above has both contextual and epistemic elements. The reference to what "a concept will pick out" under certain circumstances suggests a sort of contextual intension; but reference to how the world "turns out" suggests an epistemic element.

Chalmers also says (p. 57) that a concept's primary intension is "a function from worlds to extensions", such that "in a given world, it picks out what the referent of the concept would be if that world turned out to be actual". (The "worlds" are later refined to centered worlds.) This is similar to the characterization

above, although the use of "turned out" and "would be" arguably has a slightly different (more subjunctive, less epistemic?) flavor than the use of "turns out" and "will". Again, the referent to potential reference of a concept suggests some sort of contextual intension. This is also suggested by a later discussion (p. 60) which casts the worlds in the domain of a primary intension as Kaplanian "contexts of utterance", and which asks "how things would be if the context of the expression turned out to be W"? And again (p. 63): "the primary truth-conditions tell us how the actual world has to be for an utterance to be true in that world; that is, they specify those *contexts* in which the statement would turn out to be true.

If a primary intension is a contextual intension, what sort of contextual intension is it? The behavior that Chalmers suggests a primary intension should have is very much the sort of quasi-Fregean behavior for a 1-intension suggested initially. It seems clear that a primary intension is not intended to be an orthographic contextual intension: nothing in Chalmers' discussion suggests that in a world where 'water' means steel, the primary intension of our term 'water' picks out steel. It may be intended to be a linguistic contextual intension: footnote 21 on p. 364 suggests that Chalmers is sympathetic with the view that the word 'water' as used on Twin Earth is of the same linguistic type as ours, in which case a linguistic contextual intension may give quasi-Fregean results. It may also be that some sort of cognitive contextual intension is intended, where one holds fixed the epistemic situation of the subject. But the matter is not at all clear.

There are also a number of elements in Chalmers' discussion that suggest an epistemic understanding, however. The expression "what a concept will refer to if the worlds turns out" carries an epistemic flavor that is quite different from the subjunctive "what a concept would refer to if the world turned out"; as Kripke (1980, p. xxx) suggests, there is arguably more plausibility in the idea that it could *turn out* that 'water' refers to XYZ than that it could have *turned out* that 'water' refers to XYZ. So perhaps there is a sort of amalgam of epistemic and contextual ideas at work in this phrase.

More clearly, Chalmers' discussion of how to evaluate a primary intension has a strong epistemic element. He says

The true intension can be determined only from detailed consideration of specific scenarios: What would we say if the world turned out this way? What would we say if it turned out that way? For example, if it had turned out that the liquid in the lakes was H₂O and the liquid in the oceans was XYZ, then we would probably have said that both were water; if the stuff in oceans and lakes was a mixture of 95 percent A and 5 percent B, we would probably have said that A but not B were water.

Here, the suggestion seems to be that a term's primary intension is constituted to a speaker's or a community's dispositions to apply the term, depending on what is discovered to be the case. This suggests something at least in the vicinity of an epistemic intension. There is still a metalinguistic element in "what would we say?"; for this reason, it seems hard to extend this heuristic to such cases as evaluating "language exists" in a language-free world, and so on. But the idea of capturing the dependence of judgments about extension on discoveries about the actual world suggests something fundamentally epistemic.

Further evidence for an epistemic interpretation stems from the fact that Chalmers twice notes (notes 26 and 29, p. 366) that one can evaluate a primary intension in worlds that do not contain the original concept: he gives the example of "I am in a coma", suggesting that the primary intension should be true of centered worlds where the individual at the center is in a coma and not thinking anything. This suggests that a contextual intension is not intended (perhaps Chalmers had an extended contextual intension in mind, but it is hard to see how this could give the right result) and is compatible with an epistemic interpretation.

The strongest evidence is in footnote 21 (p. 364), where Chalmers responds to an objector who holds that 'water' on Twin Earth is a different word:

If one is worried about this ... one can think of these scenarios as *epistemic* possibilities (in a broad sense) and the conditionals as epistemic conditionals, so that worries about essential properties of words are bypassed.

Here Chalmers' response directly suggests an epistemic understanding, although he does not elaborate.[*] It should also be noted that Chalmers claims (p. 64) that a sentence is a priori when it has a necessary primary proposition (where here a proposition is an intension for a statement). And he also uses primary intensions to make an inference from conceivability to possibility (roughly, from a claim's a priori coherence to the existence of a world satisfying a claim's primary intension). These moves will be invalid in general if primary intensions are contextual intensions. But if primary intensions are epistemic intensions, it is possible that they are correct.

*[[Chalmers (1994), which develops the framework to yield an account of the narrow content of thought, says a bit more here. He says that to evaluate a primary intension, one can ask questions such as "If W turns out to be actual, what will it turn out that water is"? This is the sort of "turns-out" conditional, closely related to an indicative conditional, suggested above. He notes that the conditional is superior in a way to "If W is actual, what would the concept refer to?", since it makes clear that the concept is not required to be present in the world. Chalmers (1999) suggests that indicative conditionals can be used to evaluate primary intensions.]]

Overall, it seems that Chalmers' discussion contains elements of both a contextual and an epistemic understanding. It seems likely that what is intended is something closer to an epistemic understanding, but the details are far from clear. But in any case, it seems reasonable to suppose that Chalmers *should* have embraced an epistemic understanding, and that an epistemic approach yields the most coherent overall interpretation of his position.[*]

*[[Autobiographical remarks: I think that I tacitly intended something like an epistemic intension, but the distinction between contextual and epistemic understandings was not sufficiently clear in my mind at the time of writing, and is certainly not clear on the page. (The current paper is in part a mea culpa.) I think that if one interprets primary intensions as epistemic intensions, the uses of primary intensions elsewhere in the book are coherent. The distinction is also not entirely clear in Chalmers (1994), which was written after the relevant sections of Chalmers (1996). It is for this reason that I have not published that paper, although it has been occasionally cited; I hope to publish a revised version eventually.]]

5.6 Jackson's A-intensions

Jackson (1998) discusses "a distinction between two fundamentally different senses in which a term can be thought of as applying in various possible situations". He says:

We can think of the various situations, particulars, events, or whatever to which a term applies in two different ways, depending on whether we are considering what the term applies to under various hypotheses about which world is the actual world, or whether we are considering what the term applies to under various counterfactual hypotheses. In the first case we are considering, for each world w , what the term applies to in w , given or under the supposition that w is the actual world, our world. We can call this the A-extension of term T in world w - 'A' for actual - and call the function assigning to each world the A-extension of T in that world, the A-intension of T . In the second case, we are considering, for each world w , what T applies to in w given whatever world is in fact the actual world, and so we are, for all worlds except the actual world, considering the extension of T in a counterfactual world. We can call this the C-extension of T in w - 'C' for counterfactual - and call the function assigning to each world the C-extension of T in that world, the C-intension of T .

Here, the talk of "hypotheses about which world is the actual world", and "given or under the supposition that w is the actual world", strongly suggests that we are thinking about these worlds as a sort of epistemic possibility. One might think for a moment that talk of "what the term applies to under various hypotheses" suggests something contextual, but on reflection there is no more reason why that should be the case here than for the corresponding usage about counterfactual worlds.

Jackson does not say much more about evaluating A-intensions than this. He does say one thing that might suggest a contextual element: he says that the A-proposition (A-intension for a sentence) of 'Some water is H₂O' is contingent, because the sentence is "epistemically possible in the following sense: consistent with what is required to understand it, the sentence might have expressed something both false and discoverable to be false". The claim about what the "sentence might have expressed" strongly suggests something contextual: together with the talk of understanding, it may suggest a sort of cognitive contextual intension. But this locution is not used elsewhere.

Overall, most of Jackson's discussion is reasonably consistent with an epistemic understanding of A-intensions. Further, the purposes to which he puts the framework strongly suggests constitutive ties with apriority and the epistemic domains, and epistemic intensions work best for these purposes. If so, then A-intensions are probably best interpreted as epistemic intensions.

5.7 Kripke's epistemic duplicates

Although the work of Kripke (1980) provided the impetus for most of the two-dimensional approaches in the literature,[*] Kripke does not embrace a two-dimensional approach himself. There are numerous remarks that suggest a tacit element of two-dimensional thinking: for example, frequent remarks of the form "*Given* that ...". But there is little that formally and explicitly suggests such an approach. While it is

possible to analyze many of Kripke's epistemic claims using possible worlds, Kripke himself generally stays away from this sort of analysis.

*[[Note that Kripke 1980 was first published in 1972. One exception to this chain of influence may be Kaplan, whose work is independent of Kripke's to a significant degree.]]

There is one exception, however. Kripke notes that in cases where P is the negation of an a posteriori necessary statement, there is some intuition that "it might have turned out that P", even though P is strictly speaking impossible. For example, when a table is made of wood, there is an intuition that the table "might have turned out to be made of ice", even though that is impossible. Kripke denies the modal claim involving "might have turned out" is false in these cases, but he wants to explain it away. Similarly, Kripke notes that for statements such as 'heat is the motion of molecules', there is a sense of "apparent contingency", even though the statement is strictly necessary. Again, Kripke wants to explain this sense of contingency away.

Kripke suggests the following strategy. In these cases, although a statement is necessary, we can say that *under appropriate qualitatively identical evidential situations, an appropriate qualitatively identical statement might have been false*. And he suggests this explains the sense of apparent contingency.

What, then, does the intuition that the table might have turned out to be made of ice ... amount to? I think it means simply that there might have been *a table* looking and feeling just like this one and placed in this very position in the room, which was in fact made of ice. (p. 142)

He applies a similar strategy to 'heat is the motion of molecules', and other cases. The general principle is that when there is an intuition of apparent contingency associated with a necessary truth P, there is a qualitatively identical contingent truth P, such that P might have been false in an evidential situation qualitatively identical to the original situation. Where P is 'heat is the motion of molecules', P might be 'heat sensations are caused by the motion of molecules'.

This apparently innocuous principle packs considerable power: it enables us to reason from epistemic premises to modal conclusions. When P is "apparently contingent", or such that it seems that "it might have turned out that P", P has a distinctive *epistemic* status: to a first approximation, these claims come to the claim that P is not ruled out a priori. But the conclusion here is a *modal* one: that a certain state of affairs (involving a subject, evidence, and a statement) might really have obtained. In effect, Kripke reasons from a premise about the epistemic status of a statement to a conclusion about the possible truth of a statement token that shares a type with the original statement.

This reasoning is naturally modeled on the two-dimensional framework, understood contextually. We might say that the *evidential contextual intension* of a given statement is a function that is defined at centered worlds in which there is a subject with qualitatively identical evidence, uttering a qualitatively identical statement, and that returns the truth-value of that statement. Then Kripke is in effect suggesting that when a statement is "apparently contingent", its evidential contextual intension is contingent.

The intension in question is not completely defined, as Kripke does not define what it is for evidential situations or statements to be qualitatively identical. But it is natural to suggest that two evidential situations are identical when they are phenomenologically equivalent: that is, when what it is like to be in the first situation is the same as what it is like to be in the second. As for qualitatively identical statement: one might first suggest that this occurs when they have similar descriptive or Fregean content, but that suggestion might be inappropriate in the current context. An alternative suggestion is that two statements are qualitatively identical when they (or corresponding thoughts) play similar cognitive roles for the subject. This goes beyond Kripke and is somewhat loose, of course, but it seems at least compatible with his discussion.

Reconstructed this way, Kripke's principle takes on a familiar shape. In effect, the claim is that when a statement is apparently contingent, it has a contingent evidential contextual intension. If one substitutes aposteriority for apparent contingency and rearranges a little, one gets a familiar-looking result: if a statement has a necessary evidential contextual intension, it is a priori.

So Kripke's principle here suggests something in the vicinity of the Core Thesis for an evidential contextual intension. This should raise alarm bells. We have already seen that the Core Thesis appears to be false for any sort of contextual intension. Applying the sort of reasoning from our earlier discussion, one can straightforwardly come up with a counterexample. 'I have such-and-such evidence' is one example. For a more interesting example, let 'Bill' be a name that rigidly designates the phenomenological quality instantiated at the center of my visual field. Let us say that that quality (for me now) is phenomenal blue. Then 'Bill is phenomenal blue' is plausibly a posteriori, but it has a necessary evidential contextual intension.

This point is not simply an artifact of our reconstruction; it equally to Kripke's original claim. 'Bill is phenomenal blue' is apparently contingent in the same sort of way as paradigmatic apparently contingent statements. We have an intuition that it might have turned out that Bill was not phenomenal blue - it might have been that Bill was phenomenal red, for example. This intuition seems to be on a par with our intuitions about the table, heat, and so on. But there is no qualitatively identical evidential situation in which a qualitatively identical statement would have been false. So the Kripkean reasoning is invalid in this case.

Perhaps Kripke could shrug this off and say that the reasoning was never intended to apply in *all* cases. But this reply would have a significant cost, since the reasoning is crucial to Kripke's argument against mind-body identity theories. We have a strong intuition of apparent contingency associated with 'pain is C-fibers'; and Kripke argues that this cannot be explained away by the claim that in a qualitatively identical evidential situation, a qualitatively identical statement would have been true:

To be in the same epistemic situation that would obtain if one had a pain *is* to have a pain; to be in the same epistemic situation that would obtain in the absence of pain *is* not to have a pain. The apparent contingency of the connection between the physical state and the corresponding brain state thus cannot be explained by some sort of qualitative analog as in the case of heat. (p. 152)

Kripke uses this point to argue that as the apparent contingency cannot be explained away, the claim that 'pain is C-fibers' is not necessary at all. It follows that the claim is not true, since if it is true, it must be necessary. But it is clear that everything in the quoted passage applies equally to 'Bill'. To be in the same epistemic situation that would obtain if one had *Bill is to have Bill*; and so on. So the 'Bill' case equally cannot be explained by Kripke's paradigm. But here, one *cannot* reason from failure to satisfy Kripke's paradigm to contingency and thus to falsity: 'Bill is phenomenal blue' is clearly true and necessary. So by parity, one cannot apply this reasoning in the case of 'pain'. So Kripke's argument against the mind-body identity theory appears to be invalid.

The moral here seems to me that there is nothing special about one's *evidence* in diagnosing apparent contingency. This suggests a natural response: one might *weaken* Kripke's principle, holding simply that when a statement is apparently contingent, an appropriate qualitatively identical statement might have been false. This claim would cover the standard Kripkean cases, and would also encompass cases involving contingency of evidence itself, such as that of 'Bill'. One might think that to do this would be to kill off the argument against the identity theory, which relied crucially on qualitatively identical evidence. But there is another strand in the argument that might still apply.

Immediately after noting the considerations above, Kripke says:

The same point can be made in terms of the notion of what picks out the reference of a rigid designator. In the case of the identity of heat with molecular motion the important consideration was that although 'heat' is a rigid designator, the reference of that designator was determined by an accidental property of the referent, namely the property of producing in us sensation S. It is thus possible that a phenomenon should have been rigidly designated in the same way as a phenomenon of heat, with its reference also picked out by means of the sensation S, without that phenomenon being heat and therefore without its bring molecular motion. Pain, on the other hand, is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality. Thus pain, unlike heat, is not only rigidly designated by 'pain' but the reference of the designator is determined by an essential property of the referent.

Kripke says that this is the "same point", but in fact it is not. A close examination suggests that this point has nothing to do with evidence: it does not rely on the notion of a statement in a qualitatively identical evidential situation, but rather on that of a designator with the same manner of reference-determination. So we can apply the weakened principle suggested above, where a "qualitatively identical statement" is understood, perhaps, as a statement with the same reference-fixing intensions. Under this paradigm, the apparent contingency of 'Bill is phenomenal blue' *can* be explained away, as it is possible for a statement associated with the same reference-fixing intentions - to refer to the quality in the center of one's visual field - to be false, for example, in a case where the quality is phenomenal red. But the case of 'pain' *cannot* be explained away in these terms. So the argument against the identity theory would now seem to go through.

Nevertheless, more problems immediately arise. The reasoning here in effect invokes the claim that an

apparently contingent statement has a contingent *qualitative contextual intension*, where this sort of intension is defined at worlds centered on a statement qualitatively identical to the original, returning the referent of that token. For familiar reasons, there are counterexamples to this claim. One such is 'language exists'. Or let 'L' be a name that rigidly designates the number of actual languages. Then 'L>0' is apparently contingent: it seems that it might have turned out that L was zero, and so on. But 'L>0' has a necessary fixing contextual intension. So the principle is false. Further, it is easy to see that even the weakened Kripkean reasoning suggested above does not apply in the case of 'L>0': it is not possible for a qualitatively identical statement to be true, but the statement is necessary all the same. So even the weakened reasoning, from failure to satisfy the weaker paradigm to contingency, is invalid. So although Kripke's second argument against the identity theory is different from the first argument, it is also invalid.[*]

*[[Bealer (1996) develops Kripke's proposal using his notion of a "semantic stable" expression (one such that necessarily, in any language group in an epistemic situation qualitatively identical to ours, the expression would mean the same thing). Bealer suggests that semantically stable expressions are invulnerable to "scientific essentialism". It is not clear exactly what is required for a "qualitatively identical epistemic situation" for a language group, but it is clear that however this notion is understood, terms such as 'L' above will be semantically stable, as will rigid designators for aspects of a group's epistemic situation. So there will be a posteriori necessities involving semantically stable expressions. As in Kripke's discussion and in other cases, a broadly contextual notion (semantic stability) serves as an imperfect substitute for a broadly epistemic notion (semantic neutrality).]]

I think that the moral in both cases is that Kripke's diagnosis of intuitions about apparent contingency is incorrect: they do not turn essentially on intuitions about qualitatively identical evidential statements, and they do not turn essentially on intuitions about intuitions about qualitatively identical statements.[*] Rather, they turn on intuitions about the direct evaluation of epistemic possibilities. And we have seen that this sort of epistemic evaluation does not turn essentially on the status of possible tokens. This suggests in turn that to reason about apparent contingency, then instead of the Kripkean principles above, which tacitly involve a contextual understanding, one should appeal to principles involving an *epistemic* understanding: for example, that when a statement is apparently contingent, there is some scenario (considered as actual) in which it is false. Or, if we understand scenarios as centered worlds: that when a statement is apparently contingent, there is some centered world in which the statement's epistemic intension is true. I think that once we understand things this way, an argument against the mind-body identity theory of a sort analogous to Kripke's has a chance of succeeding. But I have written about that elsewhere.[*]

*[[There is arguably a strand in the passage above that does not turn directly on qualitatively identical statements. This is the claim that while heat is picked out by a property it has accidentally, pain is picked out by a property it has essentially. This suggests that Kripke's "same point" may come down to three different points, with potentially three different arguments in the background. I think that the third argument has the most chance of success. But I think that to make the argument work in the general case, once has to adopt something like the (highly non-Kripkean) framework of epistemic intensions.]]

*[[See e.g. Chalmers 2002p, 2002n.]]

In any case, we can see that while Kripke does not explicitly endorse a two-dimensional approach, these issues are quite close to the surface in his discussion. One can even see here an instance of a familiar two-dimensionalist pattern: trying to capture epistemic phenomena under a contextual framework, coming close, but not quite succeeding. I think that once again, the moral is that the epistemic framework is most fundamental for these purposes.

5.8 Other approaches

There are a number of other approaches that either fit within a two-dimensional framework or have something in common with these ideas. I will discuss some of these briefly.

Tichy (1984) suggests that there are two propositions corresponding to any given statement: the proposition expressed and the proposition associated. The proposition expressed by 'Phosphorus is hot' is the proposition that Venus is hot. This proposition behaves like a 2-intension. The proposition associated with 'Phosphorus is hot' is the proposition that the sentence 'Phosphorus is hot' says in English that Phosphorus is hot. This proposition behaves like a (modified) linguistic contextual intension: it is true at those worlds corresponding to a centered world where the sentence's linguistic contextual intension is true, and false at all other worlds.

Tichy uses this distinction to argue that there are no necessary a posteriori propositions. He suggests that for truths such as 'Hesperus is Phosphorus', the (trivial) proposition expressed will be necessary and a priori, while the proposition associated will be contingent and a posteriori. Whether the latter is so depends on how sentences and languages are individuated: if 'Hesperus' picks out Venus essentially, then the proposition associated will be necessary. Even if this possibility is set aside (as Tichy does), there will clearly be cases (e.g. 'language exists') of an intuitively necessary a posteriori statement with a necessary associated proposition. Tichy nowhere says that *all* intuitively a posteriori statements have an a posteriori associated proposition, and his claim that there are no necessary a posteriori propositions is arguably independent of this claim.[*] But this suggests at least that Tichy's dual-proposition account of the Kripkean phenomena does not quite get to the roots of the phenomena.

*[[Tichy sets this possibility aside on the (highly arguable) grounds that Kripke is committed to denying that it is necessary that 'Hesperus' in English picks out Venus, since on Kripke's view all that is essential to 'Hesperus' in English are linguistic conventions involving reference-fixing.]]

*[[The claim that there are no necessary a posteriori propositions is sometimes regarded (e.g. Byrne 1999) as a central aspect of the two-dimensional framework. I think it should be regarded as a further inessential claim. If one stipulates that propositions are sets of possible worlds, such a claim may be reasonable. On my own view, it is better to regard propositions as the contents of sentences (leaving open their nature) and allow that the apriority and the necessity of a statement correspond respectively to the apriority and the necessity of the proposition it expresses. If so, some propositions will be both necessary and a posteriori, and propositions will presumably themselves have a sort of two-dimensional modal structure (although they might not be reducible to this structure).]]

(Insert discussion of Perry's "reflexive content" here.)

Also relevant are some proposals for assigning content to *thought* rather than to language.[*] One such proposal is Lewis's (1979; 1986; 1994) suggestion that a subject's system of belief can be modeled by the self-attribution of a property, or equivalently, by a class of possible individuals (the subject's "doxastic alternatives"), or by a class of centered worlds. Lewis (1979) assigns this sort of content to specific beliefs, while Lewis (1986; 1994) suggests a content of this sort can be used to represent a subject's total belief system, and then in turn to show how the subject satisfies various belief ascriptions. It is clear from Lewis's discussion that this set of centered worlds behaves at least something like a 1-intension: a subject who believes "I am hungry" will have only worlds centered on hungry subjects in the set; a subject who believes "water is XYZ" will have Twin Earth centered worlds in the set, and so on.

*[[Apart from those mentioned here, some other relevant proposals concerning the narrow content of thought include the "immediate object of belief" of Brown 1986 (roughly: those propositions believed by all intrinsic duplicates), the "notional worlds" of Dennett 1982 (roughly: the set of environments in which the organism as currently constituted will flourish), the "realization conditions" of Loar 1988 (roughly: the set of worlds in which a given thought would be true if it were not a misconception), and the "notional content" of White 1991 (roughly: the class of worlds for which a subject's actions are optimal). All of these have some similarity to 1-intensions, characterized in a broadly contextual way.]]

Lewis's proposal is hard to classify directly in the present system, since he does not say much about how the relevant set of worlds (or the relevant property, or the relevant class of alternatives) is defined, apart from saying that it is determined by the subject's behavior and functional organization by the principles of belief-desire psychology (Lewis 1986, pp. 36-40).[*] Lewis says nothing to suggest that a token of specific mental or linguistic states is required at the center of the relevant worlds, however, so there is reason to think he is not invoking a contextual understanding. And Lewis's centered worlds can naturally be seen as representing a sort of epistemic possibility for the subject. So Lewis's discussion seems at least consistent with an epistemic understanding, on which the relevant set of centered worlds is a sort of epistemic intension of a subject's total belief state, consisting to a first approximation of those scenarios that verify all of a subject's beliefs.

*[[For example, the principle that a subject's actions are such that if the subject's beliefs were true, they would tend to fulfill the subject's desires: so the subject's actions constrain the subject's classes of doxastic alternatives and desire alternatives. Related here is Dennett's (1982) suggestion that a subject's "notional world" is the class of environments to which the organism as currently constituted is best fitted, and White's (1991) suggestion that a subject's "notional content" is the class of worlds for which the subjects actions are optimal, relative to available alternatives. Each of these can be seen as a suggestion for naturalizing something like a 1-intension of a subject's total belief state by appeal to the subject's behavior.]]

White (1982) sets out a multi-dimensional proposal for understanding the narrow (internally determined) content of thought and language. To simplify a little, White defines the *partial* character of a word as a function from "contexts of acquisition" (worlds centered on a functional duplicate of the original subject) to the Kaplanian character of the corresponding word in that context. This is actually a three-dimensional function (a function from centered worlds to contexts to worlds to extensions), but one can diagonalize it twice to yield a one-dimensional function: in effect, a functional contextual intension, defined at worlds centered on a functional duplicate of the original subject, returning the extension of the relevant token.

The resulting functional contextual intension will be internally determined by definition, and will be a reasonably good approximation to a Fregean semantic value. It will give anomalous results in a few cases: e.g. cognitively significant claims concerning language ('language exists') or a subject's functional organization ('I have computational structure C') may have a necessary functional contextual intension, and so will have a relatively trivial partial character. But if we assume that the epistemic intensions of a subject's tokens are determined by the subject's functional state, then the diagonalized partial character will give a reasonable approximation of an epistemic intension.

Fodor (1987) gives a related proposal for understanding narrow content. He suggests that the narrow content of a thought is a function from contexts to truth-conditions, where contexts appears to behave like worlds centered on the thought, and truth-conditions behave like 2-intensions. Like White's partial character, Fodor's function is not directly truth-evaluable (this led Fodor to eventually reject the proposal on the grounds that it is not really a sort of content), but as usual one can diagonalize it to yield a truth-evaluable content. The result is something like a conceptual contextual intension, mapping worlds centered on a token of the relevant concept or thought to that token's extension.

As with conceptual (and linguistic) contextual intensions in general, the behavior of Fodor's function will depend on how concepts and thoughts are individuated, in order to know which centered worlds are relevant. If they are syntactic mental types, then one has a sort of orthographic intension, which has uninteresting content. If they are semantic types, then it is unclear how one can specify the relevant semantic value in a noncircular way. These points (and many others) are developed by Block (1991), in a thorough critique of any proposal of this sort in accounting for narrow content.

Block makes a point worth noting here: he suggests that proposals involving a mapping from contexts "often seem to engender a *cognitive illusion* to the effect that we know *what the proposed mapping is*". I diagnose things differently. Our judgments about the mapping are grounded in perfectly reasonable intuitions about the evaluation of our terms in epistemic possibilities. The illusion on the part of these theorists is not that they grasp the mapping. Rather the illusion is that the mapping is grounded in context-dependence. Once we recognize the epistemic roots of the mapping, the problems go away.

Whether or not this diagnosis is correct, it seems fair to say that in many of the cases we have discussed, various contextual two-dimensional notions are of interest largely to the extent that they approximate epistemic two-dimensional notions. One might regard contextual notions positively as an inexpensive substitute for the epistemic notions, yielding many of the benefits without as many of the costs. Or one might regard them negatively, as a "distractor" from the more important epistemic notions that can lead to confusion because of their surface similarity. I suspect that there is something to each of these attitudes. But in any case, it seems clear that the epistemic understanding yields the deepest connections with the cognitive, epistemic, and rational domains.

Bibliography

Bealer, G. 1996. A priori knowledge and the scope of philosophy. *Philosophical Studies* 81:121-42.

- Block, N. 1991. What narrow content is not. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.
- Block, N. & Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review* 108:1-46.
- Braun, D. 1996. Demonstratives and their linguistic meanings. *Nous* 30:145-73.
- Brown, C. 1986. What is a belief state? *Midwest Studies in Philosophy* 10:357-78.
- Byrne, A. 1999. Cosmic hermeneutics. *Philosophical Perspectives*.
- Carnap, R. 1947. *Meaning and Necessity*. University of Chicago Press.
- Chalmers, D.J. 1994. The components of content. Manuscript.
<http://www.u.arizona.edu/~chalmers/papers/content.html>.
- Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D.J. 1999. Materialism and the metaphysics of modality. *Philosophy and Phenomenological Research*. <http://www.u.arizona.edu/~chalmers/papers/modality.html>
- Chalmers, D.J. 2002c. The components of content (revised version).
- Chalmers, D.J. 2002e. The nature of epistemic space.
<http://www.u.arizona.edu/~chalmers/papers/espace.html>
- Chalmers, D.J. 2002n. Consciousness and its place in nature. In (S. Stich & F. Warfield, eds) *Blackwell Guide to Philosophy of Mind*. Blackwell. <http://www.u.arizona.edu/~chalmers/papers/nature.html>
- Chalmers, D.J. 2002p. Does conceivability entail possibility? In (T. Gendler & J. Hawthorne, eds) *Imagination, Conceivability, and Possibility*. Oxford University Press.
<http://www.u.arizona.edu/~chalmers/papers/conceivability.html>
- Chalmers, D.J. 2002s. On sense and intension. <http://www.u.arizona.edu/~chalmers/papers/intension.html>
- Chalmers, D.J. 2002t. The tyranny of the subjunctive.
<http://www.u.arizona.edu/~chalmers/papers/tyranny.txt>.
- Chalmers, D.J. & Jackson, F. 2001. Conceptual analysis and reductive explanation. *Philosophical*

Review. <http://www.u.arizona.edu/~chalmers/papers/analysis.html>.

Crossley, J.N. and Humberstone, I.L. 1977. The logic of 'actually'. *Reports on mathematical logic* 8:11-29.

Davies, M. & Humberstone, I.L. 1981. Two notions of necessity. *Philosophical Studies* 58:1-30.

Dennett, D.C. 1982. Beyond belief. In (A. Woodfield, ed) *Thought and Object*. Oxford University Press.

Evans, G. 1977. Reference and contingency. *The Monist* 62:161-89.

Fodor, J.A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Frege, G. 1892. Über Sinn und Bedeutung. Translated in (P. Geach & M. Black, eds.) *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Blackwell, 1952.

Garcia-Carpintero 1998. Indexicals as token-reflexives. *Mind* 107: 529-63.

Jackson F. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford University Press.

Kaplan, D. 1979. Dthat. In (P. Cole, ed.) *Syntax and Semantics*. New York: Academic Press.

Kaplan, D. 1989. Demonstratives. In (J. Almog, J. Perry, and H. Wettstein, eds.) *Themes from Kaplan*. Oxford: Oxford University Press.

Kripke, S.A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis 1944. The modes of meaning. *Philosophy and Phenomenological Research* 4:236-50.

Lewis, D. 1979. Attitudes *de dicto* and *de se*. *Philosophical Review* 88: 513-43.

Lewis, D. 1986. *On the Plurality of Worlds*. Blackwell.

Lewis, D. 1994. Reduction of mind. In (S. Guttenplan, ed.) *Companion to the Philosophy of Mind*. Oxford: Blackwell.

Loar, B. 1988. Social content and psychological content. In (R.H. Grimm & D.D. Merrill, eds) *Contents of Thought*. University of Arizona Press.

Peacocke, C. 1992. Scenarios, concepts, and perception. In (T. Crane, ed) *The Contents of Experience*.

Cambridge University Press.

Putnam, H. 1975. The meaning of 'meaning'. In (K. Gunderson, ed.) *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press.

Rumfitt, I. 19xx. Actuality, truth, and necessary truth.

Stalnaker, R. 1978. Assertion. In (P. Cole, ed.) *Syntax and Semantics: Pragmatics, Vol. 9*. New York: Academic Press

Stalnaker 1999. *Content and Context* (foreword). Oxford University Press.

Stalnaker, R. 2001. On considering a possible world as actual.

Stalnaker, R. 1990. Narrow content. In (C.A. Anderson and J. Owens, eds.) *Propositional Attitudes*. Stanford: Center for the Study of Language and Information.

Stalnaker, R. 1991. How to do semantics for the language of thought. In (B. Loewer & G. Rey, eds.) *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

Tichy 1984. Kripke on necessary a posteriori. *Philosophical Studies*.

White, S. 1982. Partial character and the language of thought. *Pacific Philosophical Quarterly* 63:347-65.

White, S. 1991. Narrow content and narrow interpretation. In *The Unity of the Self*. MIT Press.

Yablo, S. 1999. Concepts and consciousness. *Philosophy and Phenomenological Research*.

The Components of Content

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

1 Introduction[*]

*[[This is a heavily revised version of a paper first written in 1994 and revised in 1995. Sections 1, 7, 8, and 10 are similar to the old version, but the other sections are quite different. Because the old version has been widely cited, I have made it available (in its 1995 version) at <http://consc.net/papers/content95.html>.

What follows is an application of a framework I have developed in other papers. The discussion here often passes over details that are explored in more depth in those papers. (Chalmers 2002b gives the gentlest introduction; Chalmers forthcoming a and b give full details.) The framework presented here has much in common with existing ideas in the philosophy of mind and language, especially Kaplan's (1989) and Stalnaker's (1978) two-dimensional analyses of language, Lewis's (1979) analysis of the contents of thought, and various proposals that have been made about the nature of narrow content. For some connections between these ideas, see section 9 and Chalmers (forthcoming a).]]

Here are six puzzles about the contents of thought.[*]

*[[For background material on the six puzzles, see: (1) Putnam 1975, Burge 1979; (2) Frege 1892; (3) Kripke 1979; (4) Perry 1979; (5) Schiffer 1990; (6) Kripke 1980.]]

(1) *Is content in the head?* Oscar believes that water is wet. His twin on Twin Earth, which is just like Earth except that H₂O is replaced by the superficially identical XYZ, does not. His twin's thoughts concern not water but twin water: Oscar believes that water is wet, but Twin Oscar believes that twin water is wet. This suggests that what a subject believes is not wholly determined by the internal state of the believer. Nevertheless, the cognitive similarities between Oscar and his twin are striking. Is there some wholly internal aspect of content that they share?

(2) *Frege's puzzle.* In thinking that Hesperus is Hesperus, I think about the same objects as in thinking that Hesperus is Phosphorus. But the first thought is trivial and the second is not. How can this difference in cognitive significance be reflected in a theory of content?

(3) *Kripke's puzzle.* In France, Pierre is told (in French) that London is pretty, and he believes it. Later, he arrives in London and thinks it is ugly, never suspecting that "Londres" and "London" name the same

city. It seems that Pierre simultaneously believes that London is pretty and that London is not pretty. Pierre is highly rational, however, and would never believe a contradiction. What is going on?

(4) *The problem of the essential indexical*. When I believe that I am in danger, I will take evasive action. This belief seems to be essentially indexical, or self-directed; if I merely believe that x is in danger, where (unbeknownst to me) I am x , I might do something else entirely. How can we square this indexical aspect with an account of the contents of thought?

(5) *The mode-of-presentation problem*. If Jimmy says "Lois believes that Superman can fly", he speaks truly. If he says "Lois believes that Clark Kent can fly", he speaks falsely. But on many accounts, the proposition that Clark Kent can fly is the same as the proposition that Superman can fly. If so, it seems that to believe that Clark Kent can fly, it is not enough to believe the corresponding proposition; one must believe it under an appropriate mode of presentation. What is a mode of presentation, and how can these be integrated into an account of belief ascription?

(6) *The contingent a priori*. Say it is stipulated that one meter is the length of a certain stick in Paris. Then it seems that one knows a priori that the stick is one meter long, if it exists. But it seems contingent that the stick is one meter long, as it might have been that the stick was longer or shorter than one meter. How can one have a priori knowledge of a contingent truth?

These puzzles are not unrelated. All of them suggest incompleteness in a familiar view of thought content, on which the content of a thought is tied to the external objects one is thinking about. In particular, most of them raise questions about how well such an account of thought content reflects *rational* or *cognitive* aspects of thought. Because of the dependence on external factors, this sort of content often seems to be dissociated from the rational relationships between thoughts (as witnessed by puzzles 2, 3, and 6), and from their role in guiding cognition and action (as witnessed by puzzles 1 and 4).

To resolve these and other puzzles, many have postulated a separate dimension of content — so-called "narrow content" — that depends only on the internal state of a thinker, and that is more closely tied to cognition and action.[*] The road from intuition to theory has been a difficult one, however, and no account of narrow content has yet gained widespread acceptance. It is widely held that because narrow content is internal, it lacks the sort of relation to the external world that is required to qualify as *content*. For example, many have thought that narrow content is not the sort of thing that can be true or false, as the Twin Earth cases show us that truth-conditions are not determined internally.

*[[Arguments for narrow content can be found in Dennett 1981, Fodor 1987, Lewis 1994, Loar 1988, Segal 2000, and White 1982.]]

I think that these problems are illusory, and that there is a robust and natural notion of narrow content such that narrow content has truth-conditions of its own. This can be seen by developing the idea that content has two dimensions. On the account I will give here the content of a thought can be decomposed

into two components: its *epistemic* and *subjunctive* content. Subjunctive content is a familiar external variety of content. Epistemic content has the following properties: (1) it is determined by the internal state of a cognitive system; (2) it is itself a sort of truth-conditional content; (3) it reflects the rational relations between thoughts. The first property ensures that epistemic content is a variety of narrow content. The second ensures that it is a truly semantic variety of content. The third ensures that it is central to the dynamics of cognition and action. Because of these three properties, epistemic content can help to resolve many problems in the philosophy of mind and language.

2 Intentions

In what follows, a *thought* is a token propositional attitude that aims to represent the world: for example, a belief, an expectation, or a hypothesis. Thoughts have truth-values (truth, falsity, and possibly others), and are often expressed in language by sentences. Thoughts are often (perhaps always) composed of *concepts*. Concepts are mental tokens that are often expressed in language by terms. Where thoughts have truth-values, concepts have *extensions*: for example, individuals, classes, or properties. The truth-value of a thought typically depends on the extension of the concepts involved: for example, the truth value of my thought *Hesperus is Phosphorus* depends on whether the object that is the extension of *Hesperus* is the same as the object that is the extension of *Phosphorus*.

It is a familiar idea that concepts and thoughts can be associated with an *intension*: a function from possible worlds to extensions or truth-values. The intension of a concept maps a possible world to the concept's extension in that world: in a given world, the intension of my concept *renate* picks out the class of creatures with a kidney in that world. The intension of a thought maps a world to the thought's truth-value in that world: in a given world, the intension of my thought *all renates are cordates* will be true if every creature with a kidney in that world also has a heart. In effect, a concept's intension captures the way that its extension depends on the nature of the world, and a thought's intension captures the way that its truth-value depends on the nature of the world.

It is a somewhat less familiar idea that a concept or thought can be associated with *two* intensions. First, there is an *epistemic* intension, picking out a thought or concept's extension across the space of *epistemic* possibilities. This intension captures the epistemic dependence of extension or truth-value on the way the actual world turns out. Second, there is a *subjunctive* intension, picking out a thought or concept's extension across the space of *subjunctive* or *counterfactual* possibilities. This intension captures the subjunctive dependence of extension or truth-value on counterfactual states of the world, given that the character of the actual world is already fixed. On the two-dimensional picture I will develop, a thought's epistemic intension is narrow content, while a thought's subjunctive intension is often wide content.

To give a quick illustration: for my concept *water*, the epistemic intension picks out H₂O in our world (the Earth world), and XYZ in a Twin Earth world. This reflects the fact that if I accept that my actual world is like the Twin Earth world (i.e., if I accept that the liquid in the oceans is and always has been XYZ), I should accept that water is XYZ. (drinkable liquid). By contrast, the subjunctive intension of my concept *water*, picks out H₂O in both the Earth world and the Twin Earth world. This reflects the fact

that given that water is H_2O in the actual world, the counterfactual Twin Earth world is best described as one in which water is still H_2O , and in which XYZ is merely watery stuff. As a rough approximation, we can say that the epistemic intension of *water* picks out a substance with certain superficial characteristics (e.g. a clear drinkable liquid) in any given world, while the subjunctive intension of *water* picks out H_2O in all worlds. A similar pattern exists for many other concepts; the basis of the pattern is discussed in what follows.

3 Epistemic Possibilities

Let us say that a thought is *epistemically necessary* when it can be justified a priori: that is, when there is a possible reasoning process that conclusively justifies the thought with justification independent of experience. A thought is *epistemically possible* (in a broad sense, related to but distinct from the usual philosophical sense) when it cannot be ruled out by a priori reasoning: that is, when its negation is not epistemically necessary. Intuitively, this holds when the thought does not involve an *a priori* contradiction. On this understanding, my thought *water is H_2O* is epistemically possible, as is my thought *water is XYZ*. No amount of a priori reasoning can lead to the justified rejection of either of these thoughts. For all I can know a priori, the world might be such as to make either of these thoughts true.[*]

*[[A few fine details here: (1) We can say that a thought is conclusively justified with it has the sort of justification that carries a guarantee of truth: the sort of justification carried by deduction or analysis, for example, as opposed to the non-conclusive justification carried by induction or abduction. The restriction to conclusive a priori justification is required to exclude, e.g., false mathematical thoughts that might be a priori justified by induction from true thoughts. (2) A priori indeterminate thoughts, if there are such, are neither epistemically possible nor epistemically necessary; the same goes for their negations. To handle such cases, one should say that a thought is epistemically possible when its *determinate* negation is not epistemically necessary. (3) Certain theoretical view (e.g. Salmon 1986) hold that it is knowable a priori that water is H_2O , on the grounds that 'water is H_2O ' expresses roughly the same proposition as ' H_2O is H_2O ', where this proposition is knowable a priori. Even on these controversial views, however, it is clear that the token thought *water is H_2O* is not epistemically necessary as defined above: that is, there is no reasoning process that can justify this thought a priori.]]

When a thought is epistemically possible, it is natural to hold that there are various specific *scenarios* compatible with the thought. A scenario can be thought of as a maximally specific epistemic possibility: one with all the details filled in. For example, the mere thought that water is XYZ is compatible with many epistemically possible hypotheses about the precise distribution of XYZ in my environment and about everything else that is going on in the world. Each of these maximally specific epistemically possible hypotheses corresponds to a scenario.

To flesh out this intuition further, it seems reasonable to say that some scenarios (those involving XYZ in certain distributions in my environment) *verify* my thought that water is XYZ: if I accept that the scenario obtains, I should accept that water is XYZ. Other scenarios (e.g. those involving H_2O in my environment) *falsify* my thought that water is XYZ: if I accept that the scenario obtains, I should deny that water is XYZ. Equivalently, we can say that my thought that water is XYZ *endorses* scenarios in the

first class, and *excludes* scenarios in the second class.

In effect, the space of scenarios constitutes my *epistemic space*: the space of specific epistemic possibilities that are open to me a priori. If I had no empirical beliefs, all of epistemic space would be open to me. As I acquire empirical beliefs, my epistemic space is narrowed down. Any given belief will typically *divide* epistemic space into those epistemic possibilities that it endorses and those that it excludes. The basic idea I will pursue is that the narrow content of a thought is given by the way that the thought divides epistemic space.

Scenarios are much like possible worlds. For now, we can represent scenarios using possible worlds, though I will consider potential differences later. For example, let the H₂O-world be a world like ours with H₂O in the oceans and lakes, and let the XYZ-world be a specific Twin Earth world with XYZ in the oceans and lakes. Then for all I can know a priori, my world might be qualitatively just like the H₂O world, or it might be just like the XYZ-world. So these two worlds each represent highly specific epistemic possibilities for me. We can put this by saying it is epistemically possible (in the broad sense) that the H₂O-world is actual, and that it is epistemically possible that the XYZ-world is actual.

For any given scenario, one can in principle consider the hypothesis that the scenario is actual. For any given world W (on the possible-worlds understanding), it is epistemically possible that W is actual: that is, it is epistemically possible that one's own world is qualitatively just like W. When one considers a world as an epistemic possibility in this way, one is *considering it as actual*: that is, one is thinking of it as a way one's own world may be.[*]

*[[The phrase "consider a world as actual" is due to Davies and Humberstone (1980), developing ideas presented by Evans (1979). The explication given here differs from that given by Davies and Humberstone, who do not talk explicitly about epistemic possibility, but it is in much the same spirit.]]

Any given scenario verifies some thoughts and falsifies others. Here, we can say that a scenario falsifies a thought T when the hypothesis that the scenario is actual is rationally inconsistent with T. A scenario verifies T when the hypothesis that the scenario is actual is rationally inconsistent with the negation of T. For our purposes here, it is natural to say that a thought and a hypothesis are rationally inconsistent when their conjunction is epistemically impossible: that is, when this conjunction can be ruled out by a priori reasoning. On this interpretation, a scenario verifies a thought when acceptance of the hypothesis that the scenario is actual *implies* the thought: that is, when this acceptance can lead by a priori reasoning to acceptance of the thought.

Take my thought *water is H₂O*. The H₂O-world verifies my thought: if I accept that the H₂O-world is actual, I must rationally conclude that water is H₂O. It would be rationally inconsistent to accept that the H₂O-world is actual (i.e., that the liquid surrounding me with a certain appearance and distribution is and always has been H₂O, and so on) but deny that water is H₂O. By contrast, if I accept that the XYZ-world is actual, I must rationally conclude that water is not H₂O. It would be rationally inconsistent to accept

that the XYZ-world is actual (i.e. that the liquid surrounding me with a certain appearance and distribution is and always has been XYZ, and so on), and at the same time to accept that water is H_2O . So the XYZ-world falsifies my thought *water is H_2O* , and verifies thoughts such as *water is not H_2O* , or *water is XYZ*.

There is nothing here that contradicts the claim by Kripke and Putnam that water is necessarily H_2O . Kripke and Putnam are dealing with what is often called "metaphysical" possibility and necessity, which is usually sharply distinguished from epistemic possibility and necessity. Even if it is not metaphysically possible that water is XYZ, it is epistemically possible that water is XYZ: we cannot rule out the hypothesis that water is XYZ a priori. If Kripke and Putnam are right, then when the XYZ-world is considered as a metaphysical possibility (in effect, considered as a counterfactual world different from ours), it is best described as a world where XYZ is not water. But it is clear that when it is considered as an *epistemic* possibility (i.e. considered as a way our own world may be), and when verification is defined as above, it verifies the hypothesis that water is XYZ.

The indexical character of some thoughts forces us to refine the possible-worlds understanding of scenarios. If we consider an objective world W as actual, this does not yield a fully determinate epistemic possibility. Take a world W containing both H_2O and XYZ, in the oceans and lakes of separate planets. If I consider W as actual, I am not in a position to determine whether water is H_2O or XYZ, since I do not know which planet I am on. In effect, a fully determinate hypothesis must include information about my *location* within a world. To handle this, we can represent a scenario by a *centered* world: a world marked with an individual and a time at its "center".[*] A centered world corresponds to a world *from a perspective*, marked with a viewpoint at its center. In the case above, there will be many centered worlds corresponding to W , some of which are centered on individuals on the H_2O planet, and some centered on individuals on the XYZ planet. Now, when I consider the hypothesis that a centered world W' is actual, I consider the hypothesis that my world is qualitatively just like W' , that I am the individual marked at the center of W' , and that the current time is the time marked at the center of W' . Given that sort of information in the case above, I will be in a position to determine which planet I am on, and I will be in a position to determine whether water is H_2O or XYZ.

*[[This notion is introduced by Quine (1968), who defines a centered world as a world with a marked space-time point. The definition above is due to Lewis (1979).]]

4 Epistemic Intensions

Given the above, we can naturally define a thought's *epistemic intension* as a function from scenarios to truth-values. The epistemic intension of a thought T is true at a scenario W when W verifies T , and is false at a scenario W when W falsifies T . As before, W verifies T when it is rationally inconsistent to accept that W is actual and deny T , and W falsifies T when it is rationally inconsistent to accept that W is actual and accept T .

For a more precise definition of epistemic intensions, we would need to be more precise about what it is to consider a scenario as actual. The informal understanding above suffices for many purposes, but a more detailed account can be given as follows. To consider a scenario W as actual is to consider the hypothesis that D is the case, where D is a *canonical description* of W . When scenarios are understood as centered worlds, a canonical description will conjoin an *objective* description of the character of W (including its physical and mental character, for example), with an *indexical* description of the center's location within W .[*] The objective description will be restricted to *semantically neutral* terms: roughly, terms that are not themselves vulnerable to Twin Earth thought experiments (thus excluding most names, natural kind terms, indexicals, and terms used with semantic deference). The indexical description will allow in addition indexical terms such as 'I' and 'now', to specify the center's location. We can then say that W verifies a thought T when a hypothesis that D is the case implies T .[*] Or equivalently, where S is a linguistic expression of T , W verifies T when a material conditional 'if D , then S ' is a priori. These matters are explored in much more depth in Chalmers (forthcoming a).

*[[On a full account, a canonical description of a centered world is required to be an epistemically complete statement in an idealized language that uses only semantically neutral terms and indexicals. A statement D is epistemically complete when D is epistemically possible and there is no S such that both $D \wedge S$ and $D \wedge \neg S$ are epistemically possible. A semantically neutral term is one that behaves in the same way under epistemic and subjunctive evaluation. The indexicals allowed include 'I', 'now', and any others required to characterize the center of the world. This treatment requires that for every centered world, there exists an epistemically complete description using only semantically neutral terms and indexicals. This claim can be supported by noting (i) that there will be an epistemically complete description for every world (a consequence of the idealization of the language), and (ii) that semantic non-neutrality does not in itself add expressive power in characterizing epistemic possibilities (at most, it affects the description of metaphysical possibilities). See Chalmers (forthcoming a) for more on this.]

Note that there is no requirement that a canonical description be given in a purely microphysical vocabulary or in a purely phenomenal vocabulary, or in a combination of the two. I have defended elsewhere (Chalmers 2002a, Chalmers and Jackson 2001) the claim that a conjunction ('PQTI') of microphysical, phenomenal, indexical, and "that's all" truths about the actual world implies all truths about the actual world. If so, then this sort of description will provide a canonical description of the actual world (at least if we put the microphysical description in semantically neutral form), and an analogous description will plausibly suffice for many other worlds. But this is a substantive claim, and is not built into the definition of a canonical description. If PQTI is not epistemically complete, a canonical description will need to include further information about a world. Of course there may be many canonical descriptions for a single world, but all of these will imply each other, so all will yield the same epistemic intensions.]]

In the case of *water is H_2O* , the thought's epistemic intension will be true at the H_2O -scenario (a scenario centered on Oscar surrounded by H_2O , say), and will be false at the XYZ-scenario (a scenario centered on Twin Oscar, surrounded by XYZ). On a first approximation, one might suggest that the thought's epistemic intension will be true in a scenario when the dominant clear, drinkable liquid around the center of that scenario has a certain molecular structure. This seems to capture *roughly* what it takes for us to judge that water is H_2O in the actual world, depending on how things turn out. But this sort of approximation is no replacement for the real intension. The intension itself is best evaluated by considering specific scenarios and determining the consequences for the truth-values of our thoughts.

The existence of epistemic intensions is grounded in the fact that given sufficient information about the actual world, we are in a position to know whether our thoughts are true. For example, given sufficient information about the appearance, behavior, composition, and distribution of objects and substances in my environment, I am in a position to determine whether water is H_2O . Even if the information had turned out differently, I would still have been in a position to determine whether water is H_2O . So given enough relevant information about a scenario, I am in a position to determine whether, *if* that information is correct in my own world, water is H_2O . The same goes for all sorts of other thoughts. It may be that in some cases (involving vague concepts such as *tall*, for example), a complete specification of a scenario does not settle a thought as true or false. In that case, we can say that the thought's epistemic intension is *indeterminate* at that world. But otherwise, the thought's epistemic intension will be true or false at the world.

To help evaluate an epistemic intension at a world, one can use various heuristics. One useful heuristic for evaluating the epistemic intensions of one's own thought T , expressible by a sentence S , is to intuitively evaluate an indicative conditional: 'if W is actual, is it the case that S ?' As with other indicative conditionals, one evaluates this conditional epistemically (by the "Ramsey test"): one hypothetically accepts that W is actual, and uses this to reach a rational conclusion about whether or not S is true.[*] If yes, then W verifies T ; if not, then W falsifies T . To stress the epistemic character of the conditional, one can also appeal to "turns-out" conditionals such as the following: 'if W turns out to be actual, will it turn out that S ?' For example, it seems reasonable to say that if the XYZ-world *turns out* to be actual, then it will *turn out* that water is XYZ.

*[[Note that this heuristic invokes the intuitive correctness conditions (or "assertibility conditions") of an indicative conditional, which are given by the Ramsey test, rather than the truth-conditions, whose nature is disputed. We might think of the intension defined by the indicative conditional heuristic as a "Ramsey intension". Ramsey intensions and epistemic intensions are very similar, but it is arguable that they come apart in some cases (see Yablo 2002, Chalmers 2002a). Ramsey intensions have the advantage that they do not invoke the notion of apriority, and so are available even to Quineans and others who reject this notion; this makes it clear that even a Quinean can accept a version of the general framework here. Still, my own view is that the definition in terms of apriority is more fundamental.]]

Some thoughts have a very straightforward epistemic intension. For example, it is plausible that the epistemic intension of my thought *I am a philosopher* will be true at precisely those scenarios where the individual at the center is a philosopher. The identity of the individual at the center does not matter: it might be David Chalmers, and it might be Immanuel Kant. After all, my knowledge that I am not Immanuel Kant is a posteriori, so a scenario centered on Kant represents an epistemic possibility for me in the broad sense. It seems clear that *if* I accept that the Kant scenario is my actual scenario (i.e., that I am Kant philosophizing at the center of that scenario), then I should conclude that I am a philosopher.

As for a thought such as *Hesperus is Phosphorus*: it is plausible that this thought will be verified by roughly those scenarios where the bright object visible in a certain position in the evening sky around the individual at the center is identical to the bright object visible in the morning sky around the individual at the center. Again, this captures roughly what it takes for us to judge that Hesperus is Phosphorus in the actual world, given sufficient empirical information.

With a mathematical thought such as $2+2 = 4$, or *pi is irrational*, the thought's epistemic intension will be true in all worlds. This reflects the fact that these thoughts can be justified a priori, so that the negations of these thoughts will not be rationally consistent with any a posteriori hypothesis (the conjunction will itself be epistemically impossible). The same goes even for complex mathematical thoughts whose truth we are not in a position to know ourselves. The notion of epistemic possibility and necessity involves a rational idealization away from our contingent cognitive limitations. By definition, if it is even possible for a thought to be conclusively justified a priori, then the thought is epistemically necessary. If so, the thought has a necessary epistemic intension.

(This idealization also helps to deal with a natural worry: that in practice human thinkers are too limited to entertain the complete hypothesis that a scenario *W* is actual. A thought *T* is verified by *W* if a possible thought that *W* is actual implies *T*; that is, if it is possible for a material conditional thought *if D is the case, then T* to be conclusively justified a priori, where *D* is a canonical description of *W*. This possibility may idealize away from the cognitive limitations of the thinker. Intuitively, we can think of the thinker as engaging in ideal reasoning.[*] In practice one can often avoid this sort of idealization by appealing to a relevant *partial* description *D* of *W*, where the thinker can entertain the hypothesis that *D* is the case, and where this hypothesis rationally settles the status of *T*.)

*[[As an alternative aid to the imagination, we might suppose that the thinker is assessing *T* with the aid of a supercomputer that stores the relevant information about *W* and carries out necessary a priori calculations. See Chalmers and Jackson 2001, section 4, for more on this idea.]]

It is tempting to say that conversely, when a thought has a necessary epistemic intension, it is knowable a priori. On the centered-worlds understanding of scenarios, this claim is equivalent to the thesis that when a thought is epistemically possible, it is verified by some centered world. I think this thesis is correct, and have argued for it elsewhere, but it is nontrivial, at least when the worlds in question are understood as metaphysically possible worlds. Some philosophical views entail counterexamples to this claim. For example, on some theist views it is metaphysically necessary that a god exists, but the existence of a god cannot be known a priori. If so, then *a god exists* is not a priori, but its epistemic intension will be true in all metaphysically possible worlds. In effect, there are not enough possible worlds on this view to represent all epistemic possibilities. A similar result follows from some views on which the laws of nature in our world are the laws of all worlds: there will be no worlds with different laws to represent the epistemic possibility of different laws. The same goes for some materialist views on which zombies are epistemically possible but not metaphysically possible: on some such views, no possible world will correspond to the zombie epistemic possibility.

All of these views are controversial, and I have argued elsewhere (Chalmers 2002a) that they rest on an incorrect conception of metaphysical necessity. Sometimes these views are presented as drawing support from Kripkean a posteriori necessities such as 'Hesperus is Phosphorus' and 'water is H₂O', but the Kripkean examples are all compatible with the thesis that every epistemic possibility is verified by a centered possible world. So these views require a posteriori necessities of a sort much stronger than those discussed by Kripke, and there is reason to doubt that "strong necessities" of this sort exist,

Still, one who accepts any of these views will deny the thesis that every epistemically possible thought is verified by a centered world. Such a theorist could nevertheless preserve the thesis that every epistemically possible thought is verified by a scenario, by understanding scenarios as something other than centered metaphysically possible worlds. For example, one can define a space of maximal epistemic possibilities in purely epistemic terms (perhaps using a construction from epistemically consistent thoughts or sentences), and one can then make the case that every epistemically possible thought is verified by a scenario of this sort. (For example, the theist view above entails that even if there is no godless world, there is still a godless scenario.) I have taken this purely epistemic approach to scenarios elsewhere (Chalmers forthcoming b), as it is more neutral and arguably more philosophically fundamental. For reasons of simplicity and familiarity, I will usually stay with the centered-world approach to scenarios in this paper, but it should be kept in mind that the alternative understanding is available.[*]

*[[A terminological point: I generally use "epistemic intension" for the intension defined over the space of maximal epistemic possibilities (whether or not these coincide with centered metaphysically possible worlds), while I use "primary intension" for the same sort of intension defined over the space of centered metaphysically possible worlds (whether or not these exhaust the maximal epistemic possibilities). In applications of the two-dimensional framework to metaphysical questions, the restriction to metaphysically possible worlds is often crucial, so in these contexts the issues are cast in terms of primary intensions. A central issue here is the thesis that a thought is a priori iff it has a necessary primary intension; this thesis makes a plausible but substantive claim about metaphysical possibility (one that is false if *a god exists* is metaphysically necessary but not a priori, for example). In applications of the framework to epistemic and semantic questions (such as the current discussion), the restriction to metaphysically possible worlds is less crucial, so in these contexts the issues are cast in terms of epistemic intensions. Here the claim that a thought is a priori iff it has a necessary epistemic intension can be seen as closer to definitional (if *a god exists* is necessary but not a priori, we can construe scenarios epistemically so that there will be a godless scenario). Of course if the plausible but substantive thesis above is correct, then every maximal epistemic possibility corresponds to a centered world, and primary and epistemic intensions will come to much the same thing.]]

One important note: It is tempting to suppose that the epistemic intension of a thought *T* can be evaluated in a scenario *W* by asking: what is the truth-value of *T*, as thought in *W*? But this is not so. On the present proposal, *T*'s epistemic intension can be evaluated in scenarios containing no copy of *T*; even when a copy of *T* is present, it usually plays no special role. For example, my thought *I am a philosopher* is verified by a scenario regardless of whether I think I am a philosopher there. To take a more extreme example, the epistemic intension of my thought *someone is thinking* is false in a scenario that contains no thoughts. In these cases, all that matters is the epistemic relation between the hypothesis that *W* is actual and the thought *T*. Nothing here requires that *T* be present in *W*. One might define a *different* intension (a "contextual intension"; see section 9 and Chalmers forthcoming a) using the heuristic above, but such an intension behaves in a quite different way, and will not have the same sort of epistemic properties as an epistemic intension. This will be important later.

One can define epistemic intensions for *concepts* as well as for thoughts. A concept's epistemic intension picks out its extension in a scenario considered as actual. A precise definition involves some tricky details (see Chalmers forthcoming a), so here I will simply illustrate the idea intuitively. Let us take a

singular concept *C* expressible by a term *B*. To evaluate *C*'s epistemic intension in a scenario *W*, one considers the hypothesis that *W* is actual, and uses *B* to ask: 'what is *B*?' (Here *B* is used rather than mentioned.) One can appeal to the indicative conditional 'if *W* is actual, what is *B*?'. Alternatively, one can appeal to the rational consistency of judgments of the form *C* is *such-and-such* with the hypothesis that *W* is actual.

For example, in the XYZ-scenario, the epistemic intension of my concept *water* picks out XYZ. As before, I can say: *if* the XYZ-scenario is actual, then water is XYZ. In the H₂O-scenario, on the other hand, the epistemic intension of my concept *water* picks out H₂O. More generally, one might say as a first approximation that in a given scenario, the epistemic intension of my concept *water* picks out the dominant clear, drinkable liquid found in the oceans and lakes around the individual at the center. As before, however, this is just an approximation, and the true intension corresponds to the results of considering and evaluating arbitrary scenarios as epistemic possibilities.

One can do something like this for an arbitrary concept. Even for a seemingly nondescriptive concept, such as *Gödel*, it will still be the case that given full information about a scenario and given the hypothesis that this information obtains in the actual world, one will be in a position to make a rational judgment about the identity of Gödel under that hypothesis. This mirrors the fact that given relevant information about the *actual* world, one is in a position to determine the identity of Gödel, and more generally is in a position to determine the extension of arbitrary concepts. This rational dependence of judgments about extension on information about the character of the actual world can be encapsulated in an epistemic intension.[*]

*[[For this reason, the current framework can be seen as neutral between "causal" theories of reference (on which reference is determined by a causal chain) and "descriptive" theories of reference (on which reference is determined by a description). Even a causal theorist should allow that relevant information about the actual world dictates rational judgments about our concept's extension. This methodology underlies Kripke's own arguments for the causal theory: in effect, he considers epistemic possibilities that we could discover to obtain (e.g., that a man called 'Gödel' stole the proof of the incompleteness of arithmetic from a man called 'Schmidt'), and reaches judgments about a term's extension on that basis (here, we judge that 'Gödel' will pick out the stealer, not the prover). So even on the causal theory, a term will plausibly have an epistemic intension: it is just that this epistemic intension may have a causal element. For example, for the epistemic intension of my concept *Gödel* to pick out a given individual in a scenario, it may be required that that individual stand in the right sort of causal relation to the subject at the center of the scenario. See Chalmers 2002b for more here.]]

These epistemic intensions are often difficult to characterize in independent terms, but for some concepts this characterization is straightforward. If we take a quasi-descriptive concept such as *Hesperus* (where we assume this functions to rigidly pick out the evening star in the actual world), we can say that the epistemic intension of *Hesperus* picks out the evening star around the center of an arbitrary scenario. Or if *Julius* functions to rigidly pick out the inventor of the zip, the epistemic intension of *Julius* will pick out the inventor of the zip in a given scenario.

The epistemic intension for an indexical concept is also very simple. The epistemic intension of my

concept *I* picks out the individual at the center of a scenario. The epistemic intension of *now* picks out the time at the center. The epistemic intension of *here* picks out the location of the individual at the center, at the time at the center. The epistemic intension of *today* picks out (roughly) the day that includes the time at the center. And so on.

When a thought is composed from concepts, its truth-value typically depends on the concepts' extensions. In such a case, the thought's epistemic intension will equally be determined by the concepts' epistemic intensions. For example, the epistemic intension of *A is B* will be true at a world when the epistemic intensions of *A* and *B* pick out the same individual there. One will find a similar compositionality of epistemic intensions wherever one finds compositionality of extensions.

5 Subjunctive Intensions

In contemporary philosophy, epistemic intensions are less familiar than what I will call *subjunctive intensions*. To evaluate a thought's subjunctive intension, one evaluates it in a world *considered as counterfactual*. To consider a world as counterfactual, one considers it as a *subjunctive possibility*: as a way things might have been, given that the character of the actual world is already fixed. In our world as it actually is, the liquid in the oceans and lakes is H₂O. Nevertheless, the liquid in the oceans and lakes *might have been* XYZ. So we can say that the XYZ-world *might have* obtained, and that the XYZ-world represents a subjunctive possibility.[*]

*[["Subjunctive" because this sort of possibility is grounded in the semantically subjunctive notion of what might have been the case (Kripke is explicit about this), and because the evaluation of such possibilities reflects the use of subjunctive conditionals. See Chalmers 2002b here.]]

The subjunctive intension of a thought *T* in a world *W* picks out the thought's truth-value in *W* when *W* is considered as counterfactual. Here, we grant that the character of the actual world is already fixed and ask what *would have been* the case if *W* had obtained. If *T* is expressible by a sentence *S*, we can evaluate *T*'s subjunctive intension at *W* by using *S* to ask: 'if *W* had obtained, would it have been the case that *S*?' If yes, then *T*'s subjunctive intension is true at *W*; if no, then *T*'s subjunctive intension is false at *W*. When *T*'s subjunctive intension is true at *W*, we can say that *W satisfies T*.

For example, if the XYZ-world had obtained — that is, if the liquid in the oceans and lakes *had been* XYZ — then (if Kripke and Putnam are correct) XYZ would not have been water.[*] XYZ would merely have been watery stuff, and water would still have been H₂O. If so, then the XYZ-world satisfies my thought *water is H₂O*, and the subjunctive intension of my thought is true at the XYZ-world. More generally, if Kripke and Putnam are correct, then the subjunctive intension of my thought *water is H₂O* is true at all possible worlds.

*[[I think that it is not obvious that Kripke and Putnam are correct about this, and a case can be made that it might have been that water was XYZ. For the purposes of this discussion, however, I will go along with the common view that Kripke's and Putnam's intuitions are correct. I also think that even if Kripke and Putnam are right about language, it is not

obvious that this extends to thought. But again, for the purposes of this discussion, I will go along with the common view that the modal properties of a term such as 'water' mirror modal properties of the underlying concept *water* that the term expresses.]]

It is clear that subjunctive intensions can behave quite differently from epistemic intensions. We have seen that the XYZ-world verifies *water is not H₂O*, but it satisfies *water is H₂O*. This difference is rooted in the difference between epistemic and subjunctive possibility, and the corresponding difference between considering a world as actual and as counterfactual. This is mirrored in the different behavior of indicative and subjunctive conditionals: it seems reasonable to say indicatively that if the liquid in the oceans and lakes is XYZ, then water is XYZ; but if Kripke and Putnam are right, it is not reasonable to say that if the liquid in the oceans and lakes *had been* XYZ, then water would have been XYZ. In considering a world as counterfactual, empirical facts about the actual world make a difference to how we describe it. In considering a world as actual, they do not.

Something similar goes for an indexical thought such as *I am David Chalmers*. If Kripke is right, it could not have been that I was not David Chalmers. If so, then *I am David Chalmers* is true in any world considered as counterfactual (or at least in any world where I exist). Note that there is no special need for a center in the world here: once we know all the objective facts about a counterfactual state of affairs, we know all that we need to know, even to settle indexical claims. So subjunctive possibilities can be represented by ordinary uncentered worlds, and subjunctive intensions are defined over uncentered worlds.

We can associate subjunctive intensions with concepts in a similar way. A concept's subjunctive intension picks out its extension in a world considered as counterfactual. For a concept C expressible by a term B, we can use B to ask: 'if W had been actual, what would B have been?' For example, in the case of *water*, we can say that if the XYZ-world had been actual, then water would still have been H₂O. So the subjunctive intension of *water* picks out H₂O at the XYZ-world, and plausibly picks out H₂O in all possible worlds.

For many concepts, the concept's subjunctive intension picks out its actual extension in all possible worlds. This applies in particular to *rigid* concepts: those expressible by rigid designators, such as names or indexicals, picking out the same object in all worlds. For example, Kripke argues that 'Hesperus' is a rigid designator: if Hesperus is actually Venus, then it could not have been that Hesperus was other than Venus. If so, then the subjunctive intension of *Hesperus* picks out Venus in all possible worlds. Similarly, given that 'I' is a rigid designator, the subjunctive intension of my concept *I* picks out David Chalmers in all possible worlds.

For a purely descriptive concept such as *circular* or *the inventor of the zip*, by contrast, the subjunctive intension is plausibly very similar to the epistemic intension. For example, both the epistemic and subjunctive intensions of *the inventor of the zip* plausibly pick out whoever invented the zip in a given world. Note the difference with *Julius*, which has the same epistemic intension but whose subjunctive intension picks out the actual inventor in all worlds. The difference reflects the intuition that if (for

example) Ned Kelly had invented the zip, he would have been the inventor of the zip, but he would not have been Julius. (Compare: if Ned Kelly actually invented the zip, then he is Julius.) Some concepts behave in an intermediate manner: for example, the subjunctive intension of *the discoverer of water* does not pick out the actual extension in all worlds, but it is nevertheless quite different from the epistemic intension, due to the presence of the rigid concept *water* as a constituent.

The subjunctive intension of a concept or thought always depends in some way on the concept's epistemic intension and the actual world. For a purely descriptive concept, the subjunctive intension may simply be a copy of the epistemic intension, across uncentered worlds. For a rigid concept, the subjunctive intension will correspond to the value of the epistemic intension at the actual world, projected across all possible worlds. In other cases, the dependence may be somewhat more complex, but it will still exist.

We can encapsulate this dependence by associating concepts and thoughts with a *two-dimensional intension*. This intension maps an ordered pair (V, W) consisting of a scenario and a world to an extension or a truth-value in W. When a thought T is evaluated at (V, W), it returns the truth-value of T in the counterfactual world W, under the assumption that V is actual. (If a sentence S expresses T, we can use this heuristic: 'if V is actual, then if W had obtained, would it have been the case that S?') Like an epistemic intension, a two-dimensional intension can plausibly be evaluated without relying on empirical knowledge, since all the empirical knowledge one needs is given in the first parameter V. To evaluate a thought's subjunctive intension at W, one evaluates its two-dimensional intension at (A, W), where A is the actual scenario. If we understand scenarios as centered worlds, the value of a thought's epistemic intension at a scenario W will coincide with the value of its two-dimensional intension at (W, W'), where W' is an uncentered version of W.[*] This two-dimensional intension is useful for certain purposes, but most of the time we need only appeal to a thought's epistemic and subjunctive intensions.

*[[In this way the epistemic intension can be seen as equivalent to the "diagonal" of the two-dimensional intension, in a manner reminiscent of the "diagonal proposition" of Stalnaker 1978. But see section 9 for reasons why epistemic intensions are not fundamentally diagonal intensions.]]

Within this framework, we can analyze the Kripkean "necessary a posteriori". Let us say that a sentence S is *subjunctively possible* when it is possible in the familiar Kripkean sense: that is, when a modal sentence such as 'it might have been the case that S' is true. A thought is subjunctively possible when it is expressible by a subjunctively possible sentence. Subjunctive necessity is defined correspondingly. Then it is easy to see that when a thought is subjunctively necessary, its subjunctive intension is true in all worlds, and vice versa. Cases of the Kripkean "necessary a posteriori" (e.g., *water is H₂O*) arise when a thought has a necessary subjunctive intension (the thought is true in all worlds considered as counterfactual) but a contingent epistemic intension (the thought is false in some world considered as actual). Cases of the Kripkean "contingent a priori" (e.g. *Julius invented the zip*) arise when a thought has a contingent subjunctive intension but a necessary epistemic intension.

There should be no question of whether the epistemic or the subjunctive intension is *the* intension

associated with a given concept. The full story can only be given two-dimensionally. One or the other may be more useful for various specific purposes. In matters of linguistic content across a community, the subjunctive intension often plays a central role: different users of a name or natural kind term often have quite different associated epistemic intensions while sharing the same subjunctive intension. For questions about the rational properties of thought and its role in governing action, however, we will see that the epistemic intension is central.

6 Wide and Narrow Content

Let us call a thought or concept's epistemic intension its *epistemic content*, and a thought or concept's subjunctive intension its *subjunctive content*. Let us say that when a thought or concept's content depends only on the intrinsic state of the thinker (that is, when every possible intrinsic duplicate of the thinker has a corresponding thought or concept with the same content), the content is *narrow*. Let us say that when content does not depend only on a thinker's intrinsic state (that is, when an intrinsic duplicate could have a corresponding thought or concept with different content), the content is *wide*. One can make the case that epistemic content is narrow, while subjunctive content is often wide.

It is clear that subjunctive content is often wide. For example, Oscar (on Earth) and Twin Oscar (on Twin Earth) are more or less intrinsic duplicates (abstracting away from differences due to the presence of H₂O and XYZ in their bodies), and have corresponding concepts that they express by saying 'water'. But the subjunctive intension of Oscar's concept *water* picks out H₂O in all worlds, while the subjunctive intension of Twin Oscar's concept *water* picks out XYZ in all worlds. Something similar applies to most rigid concepts, including *Hesperus* and even *I*. Here, a subjunctive intension depends on a concept's extension, which usually depends on a subject's environment, so two intrinsic duplicates can have different subjunctive intensions. In other cases, subjunctive content will not depend on the environment. For example, purely descriptive concepts such as *circular* and *the inventor of the zip* will plausibly have subjunctive intensions that are shared between duplicates. But in cases where a concept or thought's subjunctive intension depends not just on its epistemic intension but on the way the actual world turns out, we can expect that subjunctive content will be wide content.

This environment-dependence does not extend to epistemic content. A concept's epistemic content is usually quite independent of its actual extension, and of the way the actual world turns out more generally. An epistemic intension encapsulates the way in which our rational judgments about extension and truth-value depend on arbitrary empirical information; so the intension can be evaluated without knowing which epistemic possibility is actual. The factors that make subjunctive content wide content appear to be irrelevant to epistemic content.

This can be illustrated by looking at familiar cases. Take Oscar's and Twin Oscar's respective thoughts T_1 and T_2 , expressed by saying 'there is water in my pool.' Let W_1 be the Earth scenario centered on Oscar, with H₂O in the oceans and lakes and H₂O in Oscar's pool. Let W_2 be the Twin Earth scenario centered on Twin Oscar, with XYZ in the oceans and lakes and XYZ in Twin Oscar's pool. Then clearly, W_1

verifies T_1 and W_2 verifies T_2 . But also, W_2 verifies T_1 : if Oscar hypothetically accepts that W_2 is actual, he must rationally accept T_1 . Equally, W_1 verifies T_2 : if Twin Oscar hypothetically accepts that W_1 is actual, he should rationally accept T_2 . So the epistemic intensions of T_1 and T_2 are on a par with respect to these worlds.

Something similar applies to other scenarios. Let W_3 be a Twin Earth scenario centered on Twin Oscar with XYZ in the oceans and lakes, but an isolated amount of H_2O in Twin Oscar's pool. Then W_3 falsifies both T_1 and T_2 . If Oscar accepts that W_3 is actual, he should reject T_1 ; if Twin Oscar accepts that W_3 is actual, he should reject T_2 . The same goes for any other world: if W verifies T_1 , it will also verify T_2 , and vice versa. The same goes equally for any intrinsic duplicate of Oscar. We can even imagine Vat Oscar, who is a brain in a vat receiving artificial stimulation. Vat Oscar can entertain the hypothesis that his environment is just like W_1 (or W_2 or W_3) and can reach rational conclusions on that basis, and the conclusions that he reaches will mirror those of Oscar and Twin Oscar. So Vat Oscar has a thought with the same epistemic intension as Oscar's.[*] The same holds for intrinsic duplicates in general; so the epistemic content of Oscar's thought is narrow.

*[[Thus even a brain in a vat might have thoughts with epistemic content. This can be used to address Putnam's (1981) anti-skeptical argument that if he were a brain in a vat, he could not think *I am a brain in a vat*. A brain in a vat could think a thought with the appropriate epistemic content, if not the appropriate subjunctive content. It could also think a thought such as *I am in a skeptical scenario*, which has more or less identical epistemic and subjunctive content. The epistemic contents of such a thought seems sufficient to express a significant skeptical possibility, true only in worlds in which the individual at the center lacks the usual sort of epistemic contact with the surrounding world.]]

The same goes for other thoughts and concepts. For example, even though I may have a twin whose concept expressed by 'Hesperus' has a different extension and subjunctive intension, his concept nevertheless has the same epistemic intension as mine, picking out roughly the evening star near the center of any scenario. Similarly, although the *I* concepts of my twins will have an extension and subjunctive intension that differs from mine, they will have the same epistemic intension, picking out the individual at the center of any scenario.

One can even apply this analysis to the cases used by Burge (1979) to argue for the social nature of mental content. Bert has a belief that he expresses by saying 'arthritis sometimes occurs in the thighs.' In fact, arthritis is a disease of the joints and cannot occur in the thigh, so it seems that Bert has a false belief about arthritis. Twin Bert, an intrinsic duplicate of Bert, also has a belief that he expresses by saying 'arthritis sometimes occurs in the thighs'. But Twin Bert lives in a community in which the word 'arthritis' is used for a different disease, one that affects the muscles as well as the joints: we might call it 'twarthritis'. It seems that Twin Bert has a true belief about twarthritis. Where Bert believes (falsely) that he has arthritis in his thigh, Twin Bert does not: Twin Bert believes (truly) that he has twarthritis in his thigh. Burge concludes that in this sort of case, belief content is not in the head.

Here, the crucial factor is that Bert uses the term 'arthritis' with *semantic deference*, intending (at least

tacitly) to use the word for the same phenomenon for which others in the community use it. We might say that this term expresses a *deferential concept* for Bert: one whose extension depends on the way the corresponding term is used in a subject's linguistic community. It is clear that for deferential concepts, extension can depend on a subject's environment, as can subjunctive intension. The subjunctive intension of Bert's concept *arthritis* picks out arthritis in all worlds, while the subjunctive intension of Twin Bert's concept picks out twarthritis in all worlds.

Let T_1 and T_2 be the thoughts that Bert and Twin Bert express by saying 'arthritis sometimes occurs in the thighs.' Let W_1 be Bert's own centered world, with a surrounding community that uses the term 'arthritis' to refer to a disease of the joints. Let W_2 be Twin Bert's centered world, with a surrounding community that uses 'arthritis' to refer to a disease that can occur in the thigh. Then clearly W_1 falsifies T_1 and W_2 verifies T_2 . At the same time, W_2 verifies T_1 : if Bert accepts that W_2 is actual — that is, if he accepts that his linguistic community uses 'arthritis' for a disease that can occur in the thighs — then (since his concept is deferential) he should rationally accept that arthritis can occur in the thighs, and so should accept T_1 . Similarly, W_1 falsifies T_2 : if Twin Bert accepts that W_1 is actual — that is, if he accepts that his community uses 'arthritis' only for a disease of the joints — then he should reject his thought T_2 . So the epistemic intension of T_1 is false at W_1 and true at W_2 , and exactly the same is true for T_2 .

Something similar applies to any other scenarios that Bert and Twin Bert evaluate. In general, the epistemic intension of their *arthritis* concepts in those scenarios will pick out the extension of the term 'arthritis' as used in the linguistic community around the center of those scenarios. (In worlds where the term is not used, the epistemic intension will arguably be empty or indeterminate.) And the same goes for any intrinsic duplicate of Bert. Any such duplicate can entertain the hypothesis that a given scenario W is actual, and will rationally reach conclusions similar to Bert's.

One can apply the same reasoning to Putnam's case of 'elm' and 'beech', in which a subject can use the terms with different referents despite users having no substantive knowledge to differentiate the two. In this case, the terms are being used deferentially: the epistemic intension of the subject's concept *elm* picks out roughly whatever is called 'elm' around the center of a scenario, and the epistemic intension of her concept *beech* picks out roughly whatever is called 'beech' around the center of a scenario. Here again, the epistemic intension is independent of the environment. So we can see that semantic deference and "the division of linguistic labor" is quite compatible with thoughts and concepts having internally determined epistemic content.

Putnam suggests that terms such as 'water' and 'elm' show that if a concept's intension is internally determined, it cannot determine the concept's extension. The current analysis shows that this is only half-true. The epistemic intension of a concept determines its extension, and the epistemic intension is internally determined. Of course, the epistemic intension is a *centered intension*, taking a centered world as argument, at least on the possible-worlds understanding of scenarios. So Putnam's claim still holds for uncentered intensions. But any intension requires facts about the actual world to determine extension,

and it is most natural to regard the actual environment of a thinker as centered, so an internally determined centered intension is very useful here.

Why is epistemic content narrow? On the surface, this is because a thought's epistemic content is rationally prior to any knowledge of a subject's environment: it captures the way a thought's truth-value *depends* on the character of the environment, and so is independent of the environment itself. More deeply, it may be because epistemic content is defined in terms of the rational properties of thoughts, where the relevant rational properties are internally determined. For example, if one subject has a thought that is justifiable a priori, a corresponding thought in any intrinsic duplicate of that subject will also be justifiable a priori; if so, a thought's epistemic necessity is determined by the internal state of the thinker. This observation can be combined with the observation that when one subject entertains the hypothesis that a scenario W is actual, a duplicate of that subject must also be entertaining the hypothesis that W is actual. This second observation can be grounded in the fact that these hypotheses involve semantically neutral descriptions of scenarios, so there is no possibility of a "Twin Earth" difference between thinkers here. Putting these two observations together, it follows that if the hypothesis that W is actual epistemically necessitates a thought in one subject, it will also epistemically necessitate the corresponding thought in any duplicate subject. So epistemic content is narrow.

(Of course, the epistemic content of a thought will almost always depend *causally* on the external world, but it will not depend *constitutively* on the external environment. Whenever the external environment affects the epistemic contents of our thoughts, it will do so by affecting the internal state of the thinker.)

As promised, this sort of narrow content is truth-conditional. The epistemic content of a thought delivers conditions that one's actual centered world must satisfy in order for one's thought to be true. We might think of these as a thought's *epistemic* truth-conditions, as opposed to a thought's *subjunctive* truth-conditions, which govern truth across counterfactual worlds. Of course these truth-conditions can come apart at a given world: at the XYZ-world, the epistemic truth-conditions of *water is XYZ* are satisfied, but the subjunctive truth-conditions are not. This is to be expected, given the different functions of epistemic and subjunctive evaluation. One might worry that because of this, a thought could turn out to be both true and false in the actual world, but this is impossible. When evaluated at the actual world, epistemic intensions and subjunctive intensions always give the same results.

7 The Advantages of Epistemic Content

In recent times, the "content" of a thought has usually been identified with something like its subjunctive content;[*] but the epistemic content seems to be an equally good candidate. As before, there is no need to decide which is *the* content. That being said, there are a number of ways in which the epistemic content of a thought is responsible for the explanatory work that we would expect a notion of content to do.

*[[Alternatively, content is often identified with a structured proposition composed from either subjunctive intensions of the concepts involved, or from the extensions of the concepts involved (when the concepts are rigid). This sort of

structured content is more fine-grained than a subjunctive intension, but it has the same truth-conditions, and depends on the environment in a similar way. What I say below about subjunctive intensions applies equally to structured propositions. Likewise, what I say about epistemic intensions can easily be adapted to a view on which the contents of thoughts are structures composed from epistemic intensions of concepts.]]

First, epistemic content reflects the rational relations between thoughts. If one thought implies another thought a priori, the epistemic intension associated with the first entails the epistemic intension associated with the second (that is, in all scenarios in which the first intension is true, the second intension is also true). If I know that it is hot where I am now, I know that it is hot here, and vice versa; this is reflected in the fact that the epistemic contents of the two thoughts are the same. The subjunctive contents of the thoughts are very different, however. The subjunctive intension of the first thought is true at a world if it is hot where DJC is at time t in that world; the subjunctive intension of the second thought is true at a world when it is hot in place p in that world (where t and p are the time and place of the actual thoughts). The epistemic contents of the thoughts reflect their rational relationship, but the subjunctive contents do not.

It is straightforward to see why this is so. If one thought entails another a priori, then any scenario that verifies the first will verify the second. Conversely, it is plausible that if the epistemic intension of one thought entails that of another, a thinker should in principle be able to infer the second from the first by (idealized) a priori reasoning. (As before, if scenarios are identified with centered worlds and if strong necessities exist, this converse claim will be false. For example, the epistemic intension of *a god exists* might be entailed by any intension without the thought being a priori. Again, on such a view the claim can be preserved by moving to the purely epistemic understanding of scenarios.) This is not so for subjunctive intensions: entailments between these may turn on facts about the external world that are not accessible to the thinker.

This can be applied straightforwardly to explain the informativeness of a thought such as *Hesperus is Phosphorus*. Although its subjunctive intension is equivalent to that of the trivial *Hesperus is Hesperus*, its epistemic intension is quite distinct, so it is not cognitively trivial. In effect, epistemic intension here plays the role of Fregean sense. Again, it is epistemic intensions that reflect the rational properties of thoughts.

We can also invoke epistemic content in the case of Kripke's Pierre, who paradoxically seems to believe that London is pretty and that London is not pretty, without any breakdown in rationality. Pierre's concepts *Londres* and *London* have quite different epistemic intensions: in a given scenario, the first picks out (roughly) the famous city called 'Londres' that the individual at the center has heard about, whereas the second picks out (very roughly) the grimy city in which that individual has been living. The subjunctive intensions are identical, picking out London in every world. So Pierre's two beliefs *Londres is pretty* and *London is not pretty* have contradictory subjunctive intensions, but their epistemic intensions are quite compatible. Rational relations are determined by epistemic content, so contradictory subjunctive intensions support no charge of irrationality.

Intuitively, Pierre's two beliefs are rationally compatible because there are specific ways the actual world

could be that are consistent with both. There is a scenario in which 'Londres' names a faraway, beautiful city (maybe it is in India), and in which the individual at the center inhabits an entirely distinct ugly city called 'London'. For all Pierre knows and believes, such a scenario could be actual: this scenario verifies *both* of Pierre's beliefs. As long as there is such a scenario, satisfying the epistemic intensions of all of Pierre's thoughts, the *epistemic* contents of these thoughts will be compatible, and Pierre's rationality will not be in danger.

This brings out the relation between this account and Dennett's (1981) suggestion that the narrow content of a thought is reflected in the *notional world* of the thinker. We can take the notional world to be a scenario (really a class of scenarios) that verifies all of a subject's beliefs, or at least as many as possible.[*] Pierre's notional world is a world in which there is a beautiful faraway city called 'Londres', and a grimy city close at hand called 'London'. If Pierre really lived in his notional world, he would be right about everything and rarely surprised.

*[[Dennett suggests that the relevant worlds are "the environment (or class of environments) to which the organism as currently constituted is best fitted." This class may be quite different from the class of worlds that verify all of a subject's beliefs: subjects are sometimes better fitted to worlds that falsify their beliefs (when they are pessimistic or altruistic, for example); they often have beliefs about distant matters that are irrelevant to fitness; and their fitness often turns on matters about which they have no beliefs. See also the criticisms in Stalnaker 1989 and White 1991, and White's more refined account. Dennett's and White's suggestions might be seen as a first attempt at giving a naturalistic reduction of something in the vicinity of epistemic content. Such a reduction is likely to be a major project in its own right.]]

On similar grounds, one can make the case that epistemic content reflects the *cognitive* relations between thoughts. Here there is an important qualification: epistemic content as I have defined it does not distinguish the various cognitive relations that might hold between thoughts that are deductively equivalent. From the point of view of epistemic content, a complex mathematical proof is as trivial as *modus ponens*; the fine-grained cognitive dynamics of deduction lies beyond the reach of epistemic content as I have defined it here. I think a more fine-grained variety of epistemic content can handle these cases better (see Chalmers forthcoming b). I will set these issues aside here, as subjunctive content does not handle them any better, and they are largely independent of the issues at play in this paper.

A qualified thesis would be the following: insofar as epistemic content or subjunctive content reflect the cognitive relations between thoughts, the contribution of epistemic content *screens off* the contribution of subjunctive content. That is, in cases where two thoughts are cognitively related, then (1) in related cases where the epistemic content of the thoughts is held constant but the subjunctive content is varied, the cognitive relations are preserved (except insofar as cognitive relations can be affected by varying factors independent of *both* epistemic and subjunctive content, as in the deductive case); and (2) in related cases in which the subjunctive content is preserved but epistemic content is not, the cognitive relations are damaged. One can make this case straightforwardly by examining cases; the details parallel those of the discussion of the explanation of behavior, below, so I will not duplicate them here.

A third advantage of epistemic content is its suitability for a role in the explanation of behavior. It is often noted that subjunctive content seems slightly out of synchrony with what one would expect of an

explanatory psychological state. To use an example of Kaplan's (1989), if you are watching me and my pants catch fire, our respective beliefs that my pants are on fire now will have the same subjunctive content (true in all worlds in which DJC's pants are on fire at time *t*), but will lead to very different actions (I might jump into a river, while you just sit there). The difference between our actions does not seem to be something that a characterization in terms of subjunctive content alone can explain. In a similar way, belief states can produce very similar behavior for apparently systematic reasons, even when the beliefs have very different subjunctive content: witness the behavior that my twin and I produce when we think about twin water and water respectively, or the similarity between the actions of two people who think *I am hungry*. It seems that a whole dimension of the explanation of behavior is hard for subjunctive content to explain.

These explanations can be easily handled in terms of epistemic content. If you and I think *I am hungry*, the epistemic contents of our thoughts are the same, and that similarity is reflected in the similarity of our actions. When you and I both believe that my pants are on fire, on the other hand, our epistemic contents are very different, and our actions differ correspondingly. Note that this provides a straightforward solution to Perry's problem of the essential indexical: it is epistemic content, not subjunctive content, that governs action, and epistemic content, consisting in a centered intension, is a sort of *indexical* content.[*]

*[[Perry (1979) considers the possibility that centered ("relativized") propositions might provide a solution, but dismisses it on the grounds that believing that such a proposition *P* is "true for me" does not distinguish me from third parties who also believe that *P* is true for me, but act differently. The trouble is that Perry's locution "true for me" introduces an unnecessary extensional element. What distinguishes me from the third parties is rather that I believe *P simpliciter*, or better, that my belief has *P* as its epistemic content.]]

Epistemic content also accounts for the similarity of action between twin cases; this similarity reflects the fact that my beliefs about water and my twin's beliefs about twin water have the same epistemic content. But we need not move to the realm of science fiction to see the point. Two thoughts can share epistemic content even when two thinkers are quite different, as our thoughts *I am hungry* show. Even in these cases, similarities in epistemic content will lead to similarities in action, other things being equal. Suppose I think that Superman is across the road, and I want to have Superman's autograph: then other things being equal, I will cross the road.[*] If you have thoughts with similar epistemic content to mine, then you will do the same. If your thoughts share only subjunctive content with mine, while having different epistemic content — perhaps you think that Clark Kent is across the road, but want Superman's autograph — then your corresponding behavior may be quite different.

*[[To simplify the discussion, I make the happy assumption that Superman is actual and is identical to Clark Kent.]]

In general, whenever the content of a thought is causally relevant to behavior, its contribution is screened off by that of epistemic content in the following sense. If an alternative thought had the same epistemic content but different subjunctive content, the resulting behavior would have been physically indiscernible (except insofar it might be affected by changing factors independent of both sorts of content); whereas if it had the same subjunctive content but different epistemic content, the resulting behavior might have been quite different.

To see the latter point, we need only examine cases like those above. The thoughts *I am hungry* and *The guy over there is hungry* (unknowingly looking in a mirror) will lead to very different behavior, even though their subjunctive content is the same. When Lois Lane is trying to cut Clark Kent's hair, her observation "Clark's hair breaks the scissors" will prompt a reaction very different from that provoked by a corresponding thought concerning Superman. If I hear that Cary Grant is starring in a movie, I might be more likely to watch than if I hear that the movie stars Archie Leach. In all these cases, different reactions are provoked by a difference in the epistemic content of a thought. In general, whenever the epistemic content of a thought is varied, different consequences can be expected, even if subjunctive content is preserved throughout.[*] Given that epistemic content is central to cognitive relations and that cognition governs action, this is just what we would expect.

*[[Of course, thoughts like *Cary Grant is in the movie* and *Archie Leach is in the movie* might lead to the same actions despite their different epistemic content, if I know that Cary Grant is Archie Leach. But even here, there exist circumstances under which the thoughts might play a different role: imagine someone telling me that Cary Grant is not Archie Leach after all. In general, whenever two thoughts have different epistemic content, there are at least hypothetical circumstances under which the action-governing roles of the thoughts will differ.]]

By contrast, if the subjunctive content of a thought is varied but epistemic content is kept constant, behavior stays indistinguishable throughout. Perhaps, unbeknownst to me, Cary Grant is an elaborate hoax, a co-operative construction by avant-garde animators and the mass media. In such a case, my thought about Cary Grant will have no nontrivial subjunctive content, but as long as it has the same epistemic content, my behavior will be indistinguishable from the case in which he is real. Or perhaps Cary Grant is really Ludwig Wittgenstein in disguise: if so, the thought has a very different subjunctive content, but the same behavior results. Similarly, when my twin and I think *I need some more water for this pot*, the subjunctive contents of our thoughts differ, but we both go to the sink.

We can make a similar point within a single system. Take Evans' example of 'Julius', which functions to rigidly designate whoever invented the zip. Then the epistemic intensions of my concepts *Julius* and *the inventor of the zip* will be the same, but the subjunctive intensions will be very different. Despite the difference in subjunctive intensions, however, it is clear that any thoughts of the form *Julius is such-and-such* and *The inventor of the zip is such-and-such* will play very much the same role in directing cognition and action. The rigidification and consequent difference in subjunctive intension is largely irrelevant. One exception: the two concepts may behave differently in subjunctive thought, as when one judges that Julius might not have been the inventor of the zip, but not that Julius might not have been Julius. But even here the difference is accounted for by a difference in the internally determined two-dimensional intension, rather than by a difference in subjunctive content *per se*.

Some might object that there are cases in which we individuate behavior extrinsically — for example, Oscar drinks water while Twin Oscar drinks twin water — so there is a dimension of behavior that escapes epistemic content. But even in this sort of case, subjunctive content does not usually help. Even Twin Oscar, with his different subjunctive content, would drink water if he was in Oscar's present environment. What is relevant to behavior here is not subjunctive content but current environment, as we

can see by an extension of the varying-factors strategy; and I certainly do not wish to deny that current environment is relevant in the explanation of behavior.

The only cases in which there is a direct tie between subjunctive content and behavior are cases in which behavior is individuated by an intentional object: for example, Oscar searches for a glass of water, whereas Twin Oscar searches for a glass of twin water. This connection holds across all environments, as behavior only counts as water-searching if it is caused by water-thoughts. But for the same reason, this is a very weak sort of relevance for subjunctive content: as Fodor (1991) notes, in these cases the subjunctive contents of thoughts are not *causally* relevant to action, but instead are *conceptually* relevant, in effect determining the category the action falls under.[*] Subjunctive content gives us very little purchase in the *explanation* of action here, as we will only know that some behavior is water-searching if we already know that water-thoughts lie behind it. In a causal (as opposed to a conceptual) explanation of the action, epistemic content will still play the central role.

*[[See Fodor 1991 for a detailed argument along these lines. I note also that one can individuate this sort behavior intentionally but still narrowly if one individuates by epistemic content.]]

Why is epistemic content primary? To answer this question, it is useful to think of my belief contents as constituting a model of my world, a kind of map by which I steer. This is a model of the world as I find it, a centered world with me at the center, and my beliefs are constraints on that world. Beliefs constitute a model by constraining *epistemic space*: the space of epistemic possibilities that were open to me a priori. One belief rules out one group of epistemic possibilities as a candidate for the world where I am, another belief rules out another group, until only a limited class of worlds is left. I operate under the assumption that my world is one of those worlds, and if I am lucky I will not be too surprised.

My world-model is ultimately a *notional* world: a set of epistemic possibilities, such that none of these would overly surprise me if they turned out to be actual. The constraints on these possibilities are those of epistemic content. Any further constraints imposed by subjunctive content are not useful to me. The subjunctive content of my belief that the liquid in thermometers is mercury endorses only those worlds in which thermometers contain the element with atomic number x , but this constraint is so distant that if it turned out that the liquid has atomic number y , I would not be in the least surprised. In an important sense, this constraint is not reflected in my world-model at all. Insofar as my world-model is useful to me in guiding cognition and action, the constraints on it are entirely those of epistemic content.

In making a case for the primacy of epistemic content, I have not appealed to any a priori methodological principles such as the dictum that what governs behavior is in the head. The case for epistemic content has been made directly, independently of questions about physical realization. Indeed, nothing I have said implies that facts about a thinker's environment are irrelevant to the explanation of behavior. Facts about the proximal environment will clearly play an important role insofar as they affect a thinker;[*] facts about the current environment are crucial to explaining the success or failure of various actions; and facts about environmental history will at least be central to a causal explanation of a thinker's current cognitive state. All that follows from the present framework is that the environment is not relevant to the

explanation of behavior *in virtue of its role in constituting subjunctive content*. The kind of *content* that governs behavior is purely epistemic.

*[[It may even be that in certain cases, epistemic content can itself be constituted by an organism's proximal environment, in cases where the proximal environment is regarded as part of the cognitive system: if a subject's notebook is taken to be part of a subject's memory, for example (see Clark and Chalmers 1998). Here, epistemic content remains internal to a cognitive system; it is just that the skin is not a God-given boundary of a cognitive system. This is another way in which the issue between epistemic and subjunctive content runs deeper than the issue between internalism and externalism.]]

8 Belief Ascription and Psychological Explanation

All this raises a puzzle about the role of belief ascriptions in psychological explanation. If what has gone before is correct, the kind of content that governs cognition and action is epistemic content, which is narrow. But at the same time, there is strong evidence that the kind of content attributed by belief ascriptions is often wide. Does this mean that the common-sense framework of explanation of behavior in terms of belief ascription should be discarded? Alternatively, is the success of the common-sense framework evidence that something in these arguments has gone badly wrong?

Neither conclusion is justified. The present framework shows how it can at once be true that (1) belief ascriptions ascribe wide content, (2) narrow content governs action, and (3) belief ascriptions explain action. In short: belief ascriptions ascribe a combination of epistemic and subjunctive content. It is in virtue of the subjunctive component that the ascribed content is wide, and it is in virtue of the epistemic component that the ascribed content is explanatory.

A full justification of this answer requires two things. First, we need an analysis of what is attributed in belief ascriptions, so that we can see precisely what sorts of epistemic and subjunctive content are attributed. Second, we need an analysis of the role of belief ascriptions in psychological explanation, so that we can see that even in ordinary practice, it is the epistemic content attributed that carries the explanatory burden. I cannot provide anything like a complete treatment of these matters — the analysis of belief ascriptions deserves entire volumes of its own — but I can provide a preliminary sketch.

It is easy to see that ordinary belief ascriptions ascribe both epistemic and subjunctive content. If I say 'Ralph believes that Clark Kent is muscular', then in order for my utterance to be true, Ralph must have a belief that satisfies two sorts of constraints. First, the belief must have the subjunctive content of the proposition that Clark Kent is muscular (perhaps we can allow a certain amount of variation in the subjunctive content, if for example his concept of muscularity is slightly different from the norm). But that alone is not enough: a belief that Superman is muscular would have the same subjunctive content, but would not make my ascription true. As is often noted (e.g. Schiffer 1990), for the ascription to be true, the belief must involve a concept that refers to its object (Clark Kent) under an appropriate mode of presentation.[*]

*[[Some views (e.g. Salmon 1986) take ascriptions such as 'Lois believes that Clark can fly' to be strictly speaking true, so

that modes of presentation are irrelevant to truth. Even if these highly counterintuitive views are accepted, the current account can be viewed as an account of the (pragmatic) intuitive correctness conditions of belief ascriptions. Either way, we need an account of these intuitive correctness conditions to explain the function of belief ascriptions in psychological explanation.]]

In the current framework, modes of presentation are naturally seen as epistemic intensions. If Ralph refers to Clark Kent under an epistemic intension that picks out whoever is called 'Clark Kent', or one that picks out whoever is that reporter with glasses at the Daily Planet, or some more complex intension in the vicinity, my belief ascription will be true. If Ralph refers to Clark Kent under an epistemic intension that picks out the guy in the cape, or one that picks out the strongest man in the world, my belief ascription will be false. One might say that for the ascription to be true, Ralph must refer to Clark Kent under a 'Clark Kent'-appropriate epistemic intension. Here, the conditions on a 'Clark Kent'-appropriate epistemic intension are somewhat vague and unclear, and they may well be context-dependent, but it is clear from an examination of cases that they are substantive.

To take another case, if I am right in saying 'Tom believes that he is hungry', then Tom must have a belief with more or less the appropriate subjunctive content, true of all those worlds in which Tom is hungry at time *t*. But there is also a strong constraint on epistemic content. In particular, Tom must refer to himself via the epistemic intension that picks out the individual at the center in every scenario. If he sees someone in the distance clutching their belly, without realizing that he is in fact looking into a mirror, then a thought that that person is hungry has the right subjunctive content, but on the most natural reading it does not make my ascription true. The ascription will only be true if Tom's belief refers to himself under a *self*-concept, which requires a very specific sort of epistemic content. One might say that here, Tom must refer to himself under a 'he'-appropriate epistemic intension, where in context the only 'he'-appropriate epistemic intension is the purely indexical intension.

The general principle here is something like the following. A belief ascription 'x believes that S' is true when the ascriber has a belief with the subjunctive intension of S (in the mouth of the ascriber), and with an S-appropriate epistemic intension. Here, the epistemic intension is usually much less strongly constrained than the subjunctive intension. The conditions on S-appropriateness may well be complex and context-dependent; their precise nature is one of the hardest questions in the theory of belief ascriptions. Still, one can make a few generalizations. Much of the time, an epistemic intension that is not too different from the ascriber's will be S-appropriate, and much of the time, an epistemic intension that involves the terms in S itself will be S-appropriate. But this does not yield any sort of general condition. Rather, the appropriateness-conditions are best revealed by careful investigation of judgments of the ascription's truth in specific cases involving various different epistemic intensions.

In effect, this yields truth-conditions on belief ascriptions that parallel those of what Schiffer (1992) calls a "hidden-indexical" theory of belief ascription (although I have remained neutral on the ascriptions' logical form), with epistemic intensions playing the role of modes of presentation.[*] If something like this is correct, then epistemic intensions yield a solution to Schiffer's "mode of presentation" problem.[*] Epistemic intensions are perfectly suited to satisfy what Schiffer (1990, p. 252) calls "Frege's constraint" on modes of presentation: roughly, that a rational person may both believe and disbelieve that *y* is such-

and-such only if the two beliefs involve different modes of presentation of *y*. If "rationality" is interpreted to involve idealized a priori reasoning, then the satisfaction of this constraint follows from the fact that epistemic intensions reflect a priori connections between thoughts.

*[[See also Crimmins 1991 and Richard 1990. Many of the insights of these and other philosophers on the semantics of belief ascription should be straightforwardly adaptable to the present framework.]]

*[[This sort of possibility is not mentioned in Schiffer's (1990) otherwise thorough survey of potential modes of presentation.]]

We can apply this to the case of Pierre, and the ascriptions 'Pierre believes that London is pretty' and 'Pierre believes that London is not pretty.' To satisfy these ascriptions, Pierre must have beliefs with the specified subjunctive intension, referring to London under a 'London'-appropriate epistemic intension. Pierre's *London* and *Londres* concepts have different epistemic intensions, but both intensions are 'London'-appropriate. So by virtue of his belief *Londres is pretty*, Pierre satisfies the first ascription, and by virtue of his belief *London is not pretty* he satisfies the second. Before, we explained Pierre's *beliefs* by noting that his two beliefs have contradictory subjunctive intensions but compatible epistemic intensions, where only the latter is relevant to rationality. Now, we can explain the apparent contradiction in the belief *ascriptions* by noting that two different epistemic intensions can both be 'London'-appropriate, so the two ascriptions do not in fact ascribe a rational contradiction to Pierre.[*]

*[[So Kripke's "Principle of Non-Contradiction" is false: someone can rationally believe that *S* and believe that not-*S*, as long as the beliefs involve different epistemic intensions both of which satisfy the appropriate constraints.]]

We have seen that content decomposes naturally into epistemic and subjunctive content; we now see that belief ascription puts strong constraints on both. Ideally, a full theory of belief ascription will specify the nature of these constraints for any given ascription, telling us the conditions that beliefs' epistemic and subjunctive contents must satisfy in order to make the ascription true. We can think of a belief ascription as marking out a subspace in the space of ordered pairs of epistemic and subjunctive content.

Given that epistemic content governs action, it follows that if belief ascriptions are to causally explain action, it must be in virtue of the epistemic content ascribed; the subjunctive content ascribed is redundant to the explanation. To make this case properly requires examining many specific cases, but the general point can be straightforwardly illustrated. One way to see the primacy of epistemic content is to consider belief ascriptions involving empty names, such as 'Santa Claus'. These ascribe no nontrivial subjunctive content, but ascription of beliefs about Santa Claus seem to function in precisely the same way in the explanation of action as do ascriptions of beliefs about real people. We might explain Karen's agitation on Christmas Eve in terms of her belief that Santa Claus is coming, that he will not fit down the chimney, and so on. Santa's non-existence and the corresponding absence of subjunctive content make little difference to the success of such an explanation. What governs Karen's actions are her *notions* of Santa Claus; and what governs the success of the explanation is the epistemic content that these ascriptions ascribe to her. This is a *typical* case of the role of belief ascriptions in explanation: even when

non-trivial subjunctive content is ascribed (as when the referent of the name exists), it makes little difference to the patterns of explanation.

In a very wide variety of cases in which content explains action, we can see that the explanation succeeds even if the subjunctive content attributed is ignored. For instance, if we explain my opening the refrigerator in terms of my belief that there is water in the refrigerator and my desire for a glass of water, we never need to invoke the H₂O-involving subjunctive content. The explanation gains sufficient purchase from the epistemic content ascribed alone — roughly, the content that there is some of the liquid with the appropriate properties in the refrigerator, and that I want some of that liquid, and so on.[*]

*[[I leave aside here the important question of the epistemic content of desires, and the semantics of desire attributions. On my view, the epistemic content of a desire cannot in general be represented by a simple intension. Rather, it is a sort of two-dimensional intension that can endorse a different set of worlds depending on which scenario is actual. This is clearest in cases such as "I wish I were two inches taller" or "I want to be over there". The moral is that the content of desires is perhaps more deeply two-dimensional than the content of beliefs.]]

It might be objected that there are cases in which the constraints on the epistemic content ascribed by a belief ascription are weak, so that subjunctive content must be doing any explanatory work. I think that ascriptions putting weak constraints on epistemic content are rare, but let us assume they can occur: perhaps an attribution of a belief about Smith constrains the relevant epistemic intension very little.[*] Even so, if we look at explanatory practice, we see that epistemic content is still doing the real work. For example, perhaps we explain why Bev goes to the pub by saying that she wants to see Smith and believes that Smith is at the pub. Leaving aside constraints in the concepts of seeing, the pub, and so on, there is a constraint on epistemic content implicit in the 'Smith' attributions. It is implicit here that the two 'Smith' concepts in Bev's thoughts have the same epistemic intensions. If her belief and her desire had very different epistemic content associated with 'Smith' — perhaps she wants to see Batman and believes that Smith is at the pub, not knowing that Smith is Batman — the inference from those states to her action would fail. So there is a strong joint constraint on epistemic content: despite a lack of constraint on the individual beliefs, Bev is implicitly ascribed the belief that a person she wants to see is at the pub. It is *this* ascribed belief that is doing the real explanatory work, and this ascription clearly puts a heavy constraint on epistemic content. To make the case that all such examples can be similarly analyzed requires a detailed treatment, but this illustrates the general pattern.

*[[How should one analyze so-called *de re* belief attributions, of the form 'S believes of x that it is F'? In the current framework, one might adapt the proposals of Kaplan 1967 and Lewis 1979 by holding that such an ascription is true when S has a belief with the appropriate subjunctive intension, true in worlds where A has property P, where A is the referent of 'x' and P of 'F', and when the belief involves a concept that picks out A under a *de-re*-appropriate epistemic intension. Here, a *de-re*-appropriate intension is one that entails acquaintance: this requires that in any scenario in which the intension yields an extension, the subject at the center is acquainted (in the contemporary non-Russellian sense) with this extension.]]

It follows that the centrality of narrow factors in the causation of action need not overthrow the role of belief ascriptions in explaining behavior, as some (e.g. Stich 1983) have suggested it should. At most we

have shown that belief ascriptions are a somewhat rough-edged tool: they wrap both components of content into a single parcel, bringing the idle subjunctive content into play alongside the epistemic content that does all the work. But this should not surprise us; we cannot expect a folk theory to be maximally efficient.[*]

*[[Why is subjunctive content ascribed at all? I think the reasons are tied to language. First, we ascribe beliefs in the same language we use to describe the world, and when we use world-involving language to ascribe epistemic content, world-involving constraints come along naturally in the package. Second, subjunctive content is important to understanding the success of communication and of collective action. When I tell you that I have a cold, you acquire a thought whose epistemic content is different from mine, but whose subjunctive content is the same. Communication very frequently involves transmission of subjunctive content, and our collective cohesion (if not our individual actions) can often be understood in terms of shared subjunctive content. But both of these points deserve a much more extensive development.]]

In moving from common-sense psychology toward a developed cognitive science, we might expect that the kind of content that is invoked will become more purely epistemic, and that subjunctive content will be relegated to a secondary role or dropped entirely.[*] We might also expect that better tools will be developed to specify the epistemic contents of thoughts than the current rough-and-ready language of belief ascription. This might qualify as a revision of our folk notion of belief, emphasizing and refining the elements of epistemic content that are already present within it. But precisely because those elements are already present and playing a central role in our practices, such a development would fall well short of elimination.

*[[It is arguable that cognitive psychology is already mostly concerned with epistemic content rather than subjunctive content, insofar as it is concerned with content at all. For example, the psychological literature on concepts seems to be largely concerned with how concepts are applied to the actual world, concentrating on something like the epistemic intensions of the concepts involved. See Smith and Medin 1981 and Patterson 1991.]]

9 Connections and Objections

The framework outlined here is related to a number of existing proposals. There is a clear structural resemblance to other broadly two-dimensional frameworks, such as proposals by Kaplan (1989) and Stalnaker (1978) for analyzing the content of language, and proposals by White (1982) and Fodor (1987) for analyzing the contents of thought. The idea that this sort of proposal can be used to yield a sort of narrow content has been criticized by Block (1991), Stalnaker (1989; 1990), and others, and extended to an earlier version of the present proposal by Block and Stalnaker (1999). So we need to examine the relationship between these proposals, to see whether the criticisms apply. I think that on examination, the current framework differs in fundamental respects from the others, so that their problems do not arise here.

The relationship can be brought out by contrasting epistemic intensions with *contextual intensions*. A thought's contextual intension is defined by the heuristic discussed earlier: T is true in a centered world W (with T present at the center) if T is true as thought at the center of W. Likewise, the contextual intension of a concept C will return C's extension in worlds with C at the center. (One can define

contextual intensions for sentences and other linguistic expressions similarly.) There are various possible variations here: one might have different requirements for what counts as a token of T in a world, or one might require only a token of T's type (for some relevant type) rather than T itself. But however one does things, the centered worlds here are functioning as *contexts* in which a thought (or concept) occurs, and a contextual intension encapsulates the *context-dependence* of a thought's truth-value or a concept's extension.

As we saw before, contextual intensions are quite different from epistemic intensions. An obvious difference: epistemic intensions give no special role to thought tokens within a scenario, and can be evaluated in scenarios without any such tokens at the center. Thus the epistemic intension of *I am a philosopher* can be true at a scenario regardless of what the being at the center is thinking. A thought such as *someone is thinking* has an epistemic intension that is plausibly false at some centered worlds (e.g., those without any thoughts), although its contextual intension (on a natural understanding) is true at all centered worlds in which it is defined. A deeper difference: where contextual intensions represent context-dependence, with centered worlds representing contexts of thought, epistemic intensions represent *epistemic dependence*, with centered worlds representing epistemic possibilities. This is a very different conception, and yields quite different behavior. The frameworks of Kaplan and Stalnaker illustrate this. Kaplan defines the character of a linguistic expression type as a function from a context of utterance to the expression's content (roughly, subjunctive intension) relative to that context. In some ways this resembles the two-dimensional intension discussed above (in effect a function from centered worlds to subjunctive intensions), but the underlying ideas and resulting behavior are quite different. For example, on Kaplan's framework names such as 'Hesperus' and 'Phosphorus' have identical characters, picking out the same content in all contexts. This happens because the referent of a name is essential to that name, so any use of the name in any context will have the same referent. For this reason, Kaplan notes that his framework cannot provide a solution to Frege's puzzle in the case of names, natural kind terms, and the like. But as we have seen, the epistemic intension associated with a subject's use of a name behaves very differently, often picking out different individuals in different centered worlds (whether a name has its referent essentially is irrelevant on a non-contextual understanding), and holds out much more hope of dealing with Frege's puzzle.

Stalnaker defines the diagonal proposition of an expression token as a function from a world containing the token to the truth-value of the proposition that the token expresses in that world, as evaluated in that world. This bears a formal resemblance to an epistemic intension, which can be seen as equivalent to the diagonal of a two-dimensional intension. But again, the underlying ideas and resulting behavior are different. On Stalnaker's framework, the diagonal proposition of an expression is defined at worlds where it has a very different meaning. At a world where 'water is solid' means that snow is white, for example, the diagonal proposition of 'water is solid' will be true if snow is white in that world. This is quite different from an epistemic intension: if my usage is nondeferential, the use of terms such as 'water' in a scenario will be irrelevant to epistemic intensions. Stalnaker (1999) notes that because of this, diagonal propositions are not closely connected to a priori truth. This seems correct, but the problem does not generalize to epistemic intensions, which have a built-in connection to a priori truth.

White (1982) and Fodor (1987) generalize these analyses to the contents of thought. Fodor defines the

narrow content of a thought as a function from a context of thought to a thought's (wide) content in that context. White does something similar, although his account is slightly more complex and he requires that a functional duplicate of the original thinker be present in the relevant context. As before, these proposals are based on context-dependence, and give results that differ correspondingly: witness the intensions of *I am a philosopher* and *someone is thinking*.[*]

*[[Related proposals for understanding narrow content in broadly contextual terms are given by Brown 1986 and Loar 1988.]]

Block (1991) gives a number of objections to proposals of this sort. White's proposal is subject to a charge of *holism*: no two different subjects can have thoughts with the same narrow content, unless they are functional duplicates. Further, it seems that the narrow contents of a subject's thoughts all change every time the subject acquires a new belief, or indeed every time that anything happens in the mind of the subject. This problem does not apply to epistemic intensions. There is no problem with quite different thinkers having thoughts with the same epistemic intension: for example, very different people can have *I am a philosopher* thoughts with the same epistemic intension. Further, epistemic intensions will not usually change with the acquisition of new beliefs. A change in epistemic intension requires a change in a subject's rational pattern of judgments about scenarios considered as actual: a change in belief may change the subject's judgments about which scenarios are actual, but it will not usually change a subject's rational judgment about what will be the case *if* a given scenario is actual. It may be that epistemic intensions sometimes drift over time, or that corresponding thoughts of different thinkers sometimes have different epistemic intensions, but this falls well short of a general holism.

Block charges Fodor's proposal with underdetermination: it is left unclear how to evaluate the mapping across worlds. The main problem is that of "what is held constant". That is, to know which worlds fall in the domain of the intension, one needs to know just what features of the original thought must be present in the thought token at the center. If only a sort of mental syntax is held constant, the result will be an intension that delivers wildly varying results across worlds: there will be worlds where the mapping for *water* picks out steel, if a token with that mental syntax has a different meaning. If extension is held constant, then the intension will be trivial: the mapping for *water* picks out H₂O in all worlds. For better results, one might suggest that the token's *narrow content* be held constant, but that presupposes what we are trying to explain. So it seems very difficult to set things up so that the mapping yields a notion of narrow content that behaves in an appropriate way.

Again, epistemic intensions do not have this problem. There is no issues concerning what to "hold constant" across worlds here, since there is no need for the original token to be present in different worlds. Rather, we simply appeal to the *original* thought, and to its epistemic relations with the hypothesis that a given world is actual. These epistemic relations are well-defined, being grounded in the idealized rational judgments of the subject. They also do not presuppose any theoretical notion of narrow content; but they can be used to ground such a theoretical notion.

Fodor himself (1987, p. 50) raises the problem that his sort of narrow content is not semantically

evaluable (for truth and falsity), and so is not really content; rather, it is just *potential* content, delivering a content in a context. (He later rejects narrow content for this reason.) Again, epistemic content is immune to this problem. An epistemic intension is a sort of first-order content, placing direct constraints on the world, with truth-conditions of its own. Epistemic intensions can also stand in semantic relations such as entailment, and can be analyzed using semantic frameworks involving possible worlds, which allows for significant explanatory power.

Stalnaker (1990) considers the idea that some version of his diagonal proposition (or "realization conditions") might yield an account of narrow content, and raises three criticisms. First, he suggests that we cannot identity a thought independently of its content, so we cannot ask what the content of a belief *would* have been had it been a belief on twin earth. Second, he notes that diagonal propositions are defined only in worlds containing the relevant thought token, and cannot easily be extended to worlds without the thought token. Third, he notes that on this proposal narrow content is derivative on wide content (since a diagonal proposition is defined using a two-dimensional matrix which is defined using wide content), so it presupposes rather than explains wide content. In response, it is fairly clear that the first two objections apply only to contextually defined narrow content, and not to epistemically defined narrow content. On the epistemic proposal, we never need to ask what the content of a belief would have been if it had been a belief on twin earth, and narrow content is defined in a straightforward way at worlds that do not contain the relevant thought token.

In discussing his second objection, Stalnaker raises a case that is worth addressing. If Bert uses his semantically deferential concept to think *my father has arthritis in his thigh*, how can we evaluate this thought in a world in which there is no word 'arthritis' in Bert's language, and in which Bert has no thoughts about his father's health? On the epistemic framework, it is most natural to say that the epistemic intension of Bert's *arthritis* concept picks out nothing in this world. In effect, the use of a semantically deferential concept *presupposes* that one lives in a community that uses the relevant term, just as a notion such as *The present king of France* presupposes that there is a king of France. If I discover that those assumptions do not hold in my actual scenario, it is reasonable to judge that my thoughts involving these concepts lack truth-value. The same goes for alternative scenarios. In Bert's case, the epistemic intension of his thought is indeterminate in the relevant worlds. In effect, Bert's thought partitions the space of scenarios in which the background assumptions are satisfied, and says nothing about those worlds in which the assumptions are false.[*]

*[[Other concepts whose epistemic intensions have a limited domain of determinacy include perceptual demonstratives. When I think something like *That is pretty*, the referent of my demonstrative is often picked out (very roughly) as the cause of such-and-such experience. In a centered world in which there is no appropriate experience at the center, the epistemic intension may lack truth-value. This raises another subtlety: to capture the content of perceptual demonstratives, one may need to build in a "marked" experience to the center of the class of actual-world candidates, as one builds in a marked individual and time. Building this into the center will sometimes be needed to secure reference to otherwise indistinguishable experiences and their perceptual objects, as with (perhaps) a speckle in a large field, or one of the symmetrical red spots in Austin's (1990) "Two Tubes" puzzle (to which the present framework then provides a solution). In certain cases, centers may also require more than one experience, and perhaps a marked thought ("this very thought"). One might suggest that the contents of a center involve objects of "unmediated" reference: oneself, the present moment, the current thought, and perhaps certain experiences and orientations. This matter is closely connected to Russell's

suggestions about direct reference; I hope to explore it in more depth elsewhere (see also Chalmers 2002c).]]

As for Stalnaker's third objection: narrow content may be derivative on wide content on the diagonal understanding, but not on the epistemic understanding. Epistemic content can be defined quite independently of subjunctive content, and our definition of epistemic intensions makes no appeal to subjunctive evaluation. For this reason, an epistemic intension is not fundamentally a diagonal intension. *After* the fact, one can see an epistemic intension as equivalent to the diagonal of a two-dimensional intension involving both epistemic and subjunctive notions; but this complex construction is quite unnecessary to define epistemic intensions. One can characterize the first dimension of the framework in entirely epistemic terms, independently of the second dimension.

If any sort of content is derivative in the current framework, it is wide content. We have already seen that the subjunctive intension of a concept is determined by the epistemic intension in conjunction with the environment. In some cases it is a near-copy of the epistemic intension, as for simple descriptive concepts; in other cases it is determined by rigidifying the actual-world extension of the epistemic intension. By contrast, we can tell the entire story about the epistemic intension without ever involving the subjunctive intension. It therefore seems that if either intension is more fundamental, it is the epistemic intension. Still, there is no need to make too strong a claim here: both epistemic and subjunctive content are important, and both have a role to play in different domains.

Block and Stalnaker (1999) give a number of objections to the version of this framework put forward in Chalmers (1996). Many of these objections echo the objections above, and turn on interpreting the proposal via a contextual rather than an epistemic understanding.[*] Another objection is that the formal two-dimensional apparatus alone does not yield intensions with the relevant properties. This is clearly correct; but on my approach it is a substantive characterization of the intensions, not just the formal apparatus, that yields the properties in question. Block and Stalnaker also argue that the two-dimensional approach cannot explain or ground a notion of a priori truth. I have not suggested that the framework can do this; rather, I have used the notion of apriority in defining the framework. The notion of apriority, and the specific uses of it in grounding the framework, can be defended on quite independent grounds. The use of apriority in capturing the dependence of judgments about extension and truth-value on sufficient information about the world is defended at length by Chalmers and Jackson (2001).

*[[At one point, Block and Stalnaker acknowledge (in effect) that Chalmers (1996) does not intend a contextual interpretation, but suggest that a version of the "what is held constant" problem nevertheless arises in using an actual-word thought or concept to evaluate worlds without that thought or concept. I think that when things are understood in the appropriate epistemic terms, this problem clearly disappears. In fairness, it should be noted that the discussion in Chalmers (1996) is not explicit about the difference between contextual and epistemic intensions, and although the discussion tends to suggest an epistemic intension, the precise definition is left unclear. See Chalmers (forthcoming a) for discussion.]]

The current proposal also bears a resemblance to "descriptive" accounts of narrow content. It has sometimes been suggested that the narrow content of a concept such as *water* corresponds to the content of an associated description such as 'the dominant clear drinkable liquid in the environment', or some such. In response, a number of philosophers (LePore and Loewer 1986; Taylor 1989; White 1982) have

objected that even if such descriptions exist, terms such as 'liquid' are themselves susceptible to Twin Earth scenarios (e.g., where liquids are replaced by superficially identical masses of sand), so that the content of such a description is wide rather than narrow. One might think that this objection will apply to the present proposal, since I have used descriptions of this sort to characterize epistemic intensions. But importantly, the description merely provides a rough handle on the intension for the purposes of illustration. The real narrow content is a function from scenarios to extensions, and can be characterized fully only by specifying its value at specific scenarios. As soon as we move to a summarizing description in language, imperfections are introduced, and the narrowness of the content is impurified. But the intension itself remains narrow; we should not mistake the linguistic description for the real thing.[*]

*[[This might suggest that epistemic content is "ineffable". But the real problem is simply that it is difficult to capture the *epistemic content* of one expression with the *subjunctive content* of another. Just as one can capture the subjunctive content of a concept such as *water* by appealing to the equivalent subjunctive content of an expression such as 'H₂O', one might capture its epistemic content by appealing to the equivalent epistemic content of an expression such as 'the clear, drinkable liquid...'. It is hard to see why the second is any more objectionable than the first, or why it makes epistemic content any more "ineffable". Thanks to Frank Jackson for discussion on this point.]]

One would obtain a more closely related sort of "description" theory of epistemic content if one abstracted away from linguistic characterizations and regarded the relevant "descriptions" simply as properties that a referent must satisfy, or better, as relations to the thinker. If we speak merely of properties and relations, the linguistic contamination is avoided. Schiffer (1978) suggests a description theory of this sort, on which there is irreducibly *de re* reference by a thinker to himself or herself, with reference to everything else mediated by a property or relation. If we map the irreducible self-reference here to the appeal to centered worlds, and map the properties and relations to epistemic intensions, the resemblance between the accounts is clear, although Schiffer does not appeal to a two-dimensional framework, and addresses his proposal largely to the question of accounting for *de re* thought.

Another closely related idea is Lewis's (1979) proposal that belief involves the self-ascription of a property. The set of individuals satisfying a property corresponds directly to a class of centered worlds, as Lewis notes. Lewis (1994) argues that this sort of content is narrow and is primary in explanation. In effect, Lewis advocates a one-dimensional view of content, where apparent wide content is an artifact of belief ascriptions. While Lewis does not advocate understanding these contents in epistemic terms, and does not give a general characterization of the set of worlds associated with a belief, his examples suggest that these sets of worlds closely resemble those of an epistemic intension. So the present proposal appears to be highly compatible with Lewis's framework.

A residual problem for the present account is the problem of hyperintensionality. It seems that two beliefs — mathematical beliefs, for example — can have the same epistemic and subjunctive intensions, while nevertheless having intuitively different content, and playing quite different roles in cognition and action. To handle these issues, one needs a more fine-grained sort of epistemic content that goes beyond epistemic intensions as I have defined them. One might appeal to intensions over a more fine-grained space of epistemic possibilities, defined using a more demanding epistemic necessity operator that requires more than mere apriority (see Chalmers forthcoming b for some ideas here).[*] One might also

appeal to a more basic sort of content that lies behind and determines an epistemic intension. Epistemic and subjunctive intensions are aspects of the contents of thoughts, but I have not suggested that they exhaust these contents. The nature of a complete characterization of thought contents, if such a thing can be given, remains an open question.

*[[For example, one might hold that a thought is (non-ideally) epistemically necessary when it is *trivial*, in a sense to be elucidated. One could then use this notion to set up a more fine-grained epistemic space of non-ideal epistemic possibilities, and could then associate concepts and thoughts with non-ideal epistemic intensions over this space. Then two concepts or thoughts that are nontrivially a priori equivalent will have the same epistemic intension as defined in the paper, but different non-ideal epistemic intensions as defined here.]]

Another open question: is it possible to reductively explain the epistemic content of a subject's thoughts in naturalistic terms, in the way that some have attempted to explain wide content in causal or teleological terms? Certainly no such explanation is currently available. A first attempt might exploit the idea that epistemic content is mirrored in the idealized rational dispositions of the subject: perhaps a subject's actual dispositions will yield epistemic content by idealization? The normative character of the idealization may pose an obstacle to reduction, however, as will the fact that these dispositions are themselves characterized by appeal to content. My own view is that epistemic content is ultimately determined by a combination of a subject's functional organization and phenomenology.[*] If so, any attempt at explanation will need to appeal to these factors. In any case, it is arguable that if wide content depends heavily on narrow content, as the current account suggests, any adequate reductive theory of wide content will require a reductive theory of narrow content first.[*]

*[[For an argument that phenomenology is essential to the epistemic content of at least some concepts, see Chalmers 2002c. See also Horgan and Tienson (this volume, chapter 49) for arguments for phenomenally constituted narrow content that can be seen as complementing the current approach.]]

*[[Arguably, contemporary causal theories of content have been unsuccessful precisely because they attempt to account for wide content directly, without taking into account the crucial epistemic dimension involved in its determination.]]

10 Conclusion

What of the six puzzles at the start? To summarize:

(1) A thought's content decomposes into epistemic and subjunctive content, given by its epistemic and subjunctive intensions. Oscar's and Twin Oscar's thoughts differ in their subjunctive contents, and as a result ground different belief ascriptions, but their epistemic contents are the same.

(2) My thoughts that Hesperus is Hesperus and that Hesperus is Phosphorus have the same subjunctive intension but distinct epistemic intensions, as the *Hesperus* and *Phosphorus* concepts have different epistemic intensions. The triviality of the former does not imply the triviality of the latter, as it is epistemic content that governs rational relations.

- (3) Pierre's two beliefs have contradictory subjunctive intensions but compatible epistemic intensions. The apparently contradictory belief ascriptions arise because of the contradictory subjunctive intensions and because his two concepts of London have distinct epistemic intensions that can each make 'London'-involving belief ascriptions true. Rationality is governed by epistemic intensions, so there is no rational contradiction here.
- (4) The essential indexicality of belief reflects the fact that epistemic content, not subjunctive content, governs action, and that epistemic content, unlike subjunctive content, is an indexical centered intension.
- (5) The modes of presentation central to a theory of belief ascription are epistemic intensions. Belief ascriptions specify a believer's subjunctive content, and constrain the believer's epistemic content.
- (6) Instances of the contingent a priori have a necessary epistemic intension but a contingent subjunctive intension. One's cognitive world-model is constrained by epistemic content, not by subjunctive content; so a contingent subjunctive intension does not indicate a cognitive achievement.

There are many problems about the contents of thought that are not resolved by this framework. These include the problems of hyperintensionality, of a full account of belief ascriptions, and of giving a naturalistic explanation of content. Some of these matters are likely to be much more difficult than the puzzles at issue in this paper, but the two-dimensional approach at least clarifies the lay of the land.

Acknowledgments

Thanks to too many people to mention, but especially Ned Block, Curtis Brown, Frank Jackson, David Lewis, Mark Sainsbury, Robert Stalnaker, and two reviewers. Thanks also to audiences at talks between 1994 and 1997 at Arizona, Cornell, Memphis, Princeton, Rice, UCLA, UC Santa Barbara, UC Santa Cruz, Washington University, Yale, and the Australasian Association of Philosophy. For comments on the revised version, thanks to Torin Alter, Chris Evans, Brie Gertler, and Daniel Stoljar.

References

- Austin, D. F. 1990. *What's the Meaning of "This"?* Cornell University Press.
- Block, N. 1991. What narrow content is not. In B. Loewer and G. Rey, eds., *Meaning in Mind: Fodor and His Critics*. Blackwell.
- Block, N. and Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review* 108:1-46.
- Brown, C. 1986. What is a belief state? *Midwest Studies in Philosophy* 10:357-78.

- Burge, T. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4: 73-122.
- Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D. J. 2002a. Does conceivability entail possibility? In T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford University Press. <http://consc.net/papers/conceivability.html>.
- Chalmers, D. J. 2002b. On sense and intension. *Philosophical Perspectives* 16. <http://consc.net/papers/intension.html>.
- Chalmers, D. J. 2002c. The content and epistemology of phenomenal belief. In Q. Smith and A. Jokic (eds.), *Consciousness: New Philosophical Essays*. Oxford University Press. <http://consc.net/papers/belief.html>.
- Chalmers, D. J. (forthcoming a). The foundations of two-dimensional semantics. <http://consc.net/papers/foundations.html>.
- Chalmers, D. J. (forthcoming b). The nature of epistemic space. <http://consc.net/papers/espace.html>.
- Chalmers, D. J. and Jackson, F. 2001. Conceptual analysis and reductive explanation. *Philosophical Review* 110:315-61. <http://consc.net/papers/analysis.html>.
- Clark, A. and Chalmers, D. J. 1998. The extended mind. *Analysis* 58:7-19. <http://consc.net/papers/extended.html>.
- Crimmins, M. 1991. *Talk about Beliefs*. MIT Press.
- Davies, M. K. and Humberstone, I. L. 1980. Two notions of necessity. *Philosophical Studies* 38:1-30.
- Dennett, D.C. 1981. Beyond belief. In A. Woodfield (ed.), *Thought and Object*. Oxford University Press.
- Evans, G. 1979. Reference and contingency. *The Monist* 62:161-89.
- Fodor, J. 1987. *Psychosemantics*. MIT Press.
- Fodor, J. 1991. A modal argument for narrow content. *Journal of Philosophy* 88:5-26.
- Frege, G. 1892. Über Sinn und Bedeutung. Translated in (P. Geach and M. Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*. Blackwell, 1952.

Horgan, T. and Tienson, J. (this volume). The intentionality of phenomenology and the phenomenology of intentionality.

Kaplan, D. 1967. Quantifying in. *Synthese* 19:178-214.

Kaplan, D. 1989. Demonstratives. In J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*. Oxford University Press.

Kripke, S. A. 1979. A puzzle about belief. In A. Margalit (ed.), *Meaning and Use*. D. Reidel.

Kripke, S. A. 1980. *Naming and Necessity*. Harvard University Press.

LePore, E. and Loewer, B. 1986. Solipsistic semantics. *Midwest Studies in Philosophy* 10:595-614.

Lewis, D. 1979. Attitudes *de dicto* and *de se*. *Philosophical Review* 88:513-43.

Lewis, D. 1994. Reduction of mind. In S. Guttenplan (ed.), *Companion to the Philosophy of Mind*. Blackwell.

Loar, B. 1988. Social content and psychological content. In R.H. Grimm and D.D. Merrill (eds.), *Contents of Thought*. University of Arizona Press.

Patterson, S. 1991. Individualism and semantic development. *Philosophy of Science* 58:15-35.

Perry, J. 1977. Frege on demonstratives. *Philosophical Review* 86:474-97.

Perry, J. 1979. The problem of the essential indexical. *Nous* 13:3-21

Putnam, H. 1975. The meaning of 'meaning'. In K. Gunderson (ed.), *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press.

Putnam, H. 1981. *Reason, Truth, and History*. Cambridge University Press.

Quine, W. V. 1968. Propositional objects. *Critica* 2(5):3-22.

Richard, M. 1990. *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*. Cambridge University Press.

Salmon, N. 1986. *Frege's Puzzle*. MIT Press.

Schiffer, S. 1978. The basis of reference. *Erkenntnis* 13:171-206.

- Schiffer, S. 1990. The mode-of-presentation problem. In C.A. Anderson and J. Owens (eds.), *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*. CSLI Press.
- Schiffer, S. 1992. Belief ascription. *Journal of Philosophy* 87:602-14.
- Segal, G. 2000. *A Slim Book about Narrow Content*. MIT Press.
- Smith, E. E. and Medin, D. L. 1981. *Categories and Concepts*. Harvard University Press.
- Stalnaker, R. 1978. Assertion. In P. Cole (ed.), *Syntax and Semantics: Pragmatics, Vol. 9*. Academic Press
- Stalnaker, R. 1989. On what's in the head. *Philosophical Perspectives* 3:287-316.
- Stalnaker, R. 1990. Narrow content. In C.A. Anderson and J. Owens (eds.), *Propositional Attitudes*. Stanford: Center for the Study of Language and Information.
- Stalnaker, R. 1999. *Context and Content*. Oxford University Press.
- Stich, S. 1983. *From Folk Psychology to Cognitive Science*. MIT Press.
- Taylor, K. 1989. Narrow content functionalism and the mind-body problem. *Nous* 23:355-72
- White, S. 1982. Partial character and the language of thought. *Pacific Philosophical Quarterly* 63:347-65.
- White, S. 1991. Narrow content and narrow interpretation. In *The Unity of the Self*. Cambridge, MA: MIT Press.
- Yablo, S. 2002. Coulda, woulda, shoulda. In T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford University Press.

The Components of Content

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

1 Introduction

Here are six puzzles about the contents of thought.[*]

*[[For background material on the six puzzles, see: (1) Putnam 1975, Burge 1979; (2) Frege 1892; (3) Kripke 1979; (4) Perry 1979; (5) Schiffer 1990; (6) Kripke 1980.]]

(1) *Is content in the head?* I believe that water is wet. My twin on Twin Earth, which is just like Earth except that H₂O is replaced by the superficially identical XYZ, does not. His thoughts concern not water but twin water: I believe that water is wet, but he believes that twin water is wet. It follows that that what a subject believes is not wholly determined by the internal state of the believer. Nevertheless, the cognitive similarities between me and my twin are striking. Is there some wholly internal aspect of content that we might share?

(2) *Frege's puzzle.* The thought that Hesperus is Hesperus seems to express the same proposition as the thought that Hesperus is Phosphorus, but the former is trivial and the latter is not. How can this difference in cognitive significance be reflected in a theory of semantic content?

(3) *Kripke's puzzle.* In France, Pierre is told that "Londres est jolie", and he believes it. Later, he arrives in London and thinks it is ugly, never suspecting that "London" and "Londres" name the same city. It seems that Pierre simultaneously believes that London is pretty and that London is not pretty. But Pierre is highly rational, and would never believe a contradiction. What is going on?

(4) *The problem of the essential indexical*. When I believe that I am in danger, I will take evasive action. This belief state seems to be essentially indexical, or self-directed; if I merely believe that x is in danger, where x happens to be me, I might do something else entirely. How can we square this indexical aspect with a propositional account of the contents of thought?

(5) *The mode-of-presentation problem*. If Jimmy says "Lois believes that Superman can fly", he speaks truly. If he says "Lois believes that Clark Kent can fly", he speaks falsely. But on many accounts, the proposition that Clark Kent can fly is the same as the proposition that Superman can fly. If so, it seems that to believe that Clark Kent can fly, it is not enough to believe in the corresponding proposition; one must believe it under an appropriate mode of presentation. But what is a mode of presentation, and how can these be integrated into an account of belief ascription?

(6) *The contingent a priori*. Say it is stipulated that one meter is the length of a certain stick in Paris. Then one knows *a priori* that the stick is one meter long, if it exists. But the corresponding proposition is contingent, as the stick might have been longer or shorter. How can one have *a priori* knowledge of the truth of a contingent proposition?

These puzzles are not unrelated. All of them suggest an incompleteness in a propositional view of thought content, on which the content of a belief that P is identified with the conditions under which P is true, or with some other version of the content of the proposition that P . In particular, most of them raise questions about how well a propositional account of thought content reflects *rational* or *cognitive* aspects of thought. Because a thought's propositional content is strongly tied to the external objects of the thought, a propositional account seems to have trouble capturing the rational relationships between thoughts (as witnessed by puzzles 2, 3, and 6), and their role in guiding cognition and action (as witnessed by puzzles 1 and 4).

To resolve these and other puzzles, many have postulated a separate internal dimension of content - so-called "narrow content" - that is more closely tied to cognition and action.[*] But the road from intuition to theory has been a difficult one, and no account of narrow content has yet gained widespread acceptance. Perhaps the greatest difficulty is that because narrow content is internal, it seems to lack the sort of relation to the external world that is required to qualify as *content*. Many have thought that narrow content is not the sort of thing that can be true or false, for example, as the Twin Earth cases show us that truth-conditions are not determined internally.[*]

*[[Arguments for narrow content can be found in Fodor 1987, Lewis 1994, Loar 1988, and White 1982.]]

*[[Even supporters of narrow content have often conceded this point: e.g. Fodor 1987, Block 1988.]]

I think that these problems are illusory, and that there is a robust and natural notion of narrow content such that narrow content has truth-conditions of its own. This can be seen by developing the insight, implicit in much contemporary work in the philosophy of mind and language, that content is two-dimensional. It turns out that the content of a thought can be decomposed into two components: its *notional* and *relational* content. Relational content is the familiar external variety of propositional content. Notional content is the additional component, with the following properties: (1) it is determined by the internal state of a cognitive system; (2) it is itself a sort of propositional content, with its own truth-conditions; (3) it governs the rational relations between thoughts. The first property ensures that notional content is a variety of narrow content. The second ensures that it is a truly semantic variety of content. The third ensures that it is central to the dynamics of cognition and action. These three properties together, placed within the two-dimensional framework, allow many problems in the philosophy of mind and language, including the six puzzles above, to be straightforwardly resolved.[*]

*[[The framework I outline was suggested by a reading of Kripke 1980 and Putnam 1975, but it has much in common with other existing ideas in the philosophy of mind and language that I will point to along the way. Most obviously there is a clear relationship to the two-dimensional accounts of the semantics and pragmatics of language given by Kaplan (1989) and Stalnaker (1978), but there are deep links to many other ideas besides. I hope that it is a virtue of the present proposal that it suggests direct links between ideas that might at first glance seem only distantly related.]]

2 Primary and Secondary Intensions

In what follows, an *intension* is a function over the space of possible worlds, taking a world as argument and returning an entity in that world. A *centered world* is a possible world with a marked individual and time (the "center" of that world).[*] W and W^* are the spaces of possible worlds and centered worlds respectively, and R is the space of entities in those worlds. A *concept*, pre-theoretically, is an element of a specific thought, of the sort that is an appropriate candidate for reference - that is, it either refers or has empty reference. (Note that a concept is a token rather than a type, tied to a specific thinker on a specific occasion.) A concept, in thought, is the analog of a term in language. For ease of exposition I will restrict attention to concepts that are associated with terms in the thinker's language ("I", "water", "arthritis"), but the account can be easily extended to concepts without linguistic counterparts.

*[[This notion is introduced by Quine (1968), who defines a centered world as a world with a marked space-time point. The definition above is due to Lewis (1979).]]

It is a familiar idea that every concept has an associated intension, mapping a possible world to the referent of the concept within that world. It is less commonly acknowledged that there are in fact two intensions associated with a concept, although this point is implicit in much recent work on meaning and reference. It is central to the work of Kripke and Putnam, for example, that there are two ways in which the reference of a concept in a world can depend on that world. First there is a dependency whereby reference is fixed in the actual world, depending on how the world turns out. My concept "water" refers to H₂O given that the world has turned out one way, but if it had turned out differently my concept might have referred to XYZ. Second, there is a dependency whereby reference in counterfactual worlds is

determined, given that the actual world is held constant. For example, given that the actual world turns out as it has, my concept "water" picks out H₂O in all counterfactual worlds. Corresponding to these two dependencies are two intensions, a concept's *primary* and *secondary* intensions.

The *primary intension* of a concept is a function $f_1: W^* \rightarrow R$ from centered worlds to referents. At a centered world w , the primary intension returns the referent of the concept in w when w is *considered as actual*[*] - that is, when it is considered as a candidate for the actual world of the thinker. This captures the way the reference of the concept is fixed, depending on how the actual world turns out - that is, depending on which centered world turns out to be the thinker's actual world. We can think of the primary intension as encapsulating what it takes for an entity in the actual world to qualify as the concept's referent.

*[[This phrase is due to Davies and Humberstone (1980). It is a familiar idea from two-dimensional modal logic that statements about possible worlds can be evaluated differently depending on whether the worlds are considered as actual or considered as counterfactual. The basic idea is presented in Evans 1979.]]

To evaluate a concept's primary intension at a centered world w , we can ask as a heuristic: to what would the concept refer, if w turned out to be actual? For example, if the actual world turned out to contain XYZ as the dominant clear liquid in the oceans and lakes on my planet, then my concept "water" would refer to XYZ. So the primary intension maps the XYZ-world to XYZ. On the other hand, given that the actual world turns out as it has, with H₂O in the oceans and lakes, my concept "water" refers to H₂O. So the primary intension maps the H₂O-world (the *real* actual world[*]) to H₂O.

*[[This is another construction familiar from two-dimensional modal logic (Davies and Humberstone, p. 10), to make explicit the fact that of the various worlds that might turn out to be actual, one of them *is* actual.]]

An alternative heuristic for determining the primary intensions of one's own concepts (such as "water") is to ask questions analogous to: if w turned out to be actual, what would it turn out that water is? There is a clear sense of the "turns out" locution on which it is reasonable to say that if the XYZ-world turned out to be actual, it would turn out that water is XYZ; this is the sense that is relevant in the heuristic. (On this reading, "it would turn out that" creates an intensional context even for a rigid designator.) I avoid this locution for the most part, as it can be interpreted in different ways,[*] but once the ambiguity is resolved this heuristic seems quite clear.[*] (In one way it may be better than the heuristic above, as it makes it clearer that the concept itself need not be present in w , as discussed below.)

*[[Sometimes philosophers resist this intensional reading of the "turns out" locution under the influence of Kripke (1980), but the commonsense reading remains viable even once Kripke's main point is granted. Kripke himself (p. 143) allows a sense in which gold could turn out not to have atomic number 79, although he insists that gold could not *have turned out* not to have atomic number 79. (To my ear, "could have turned out that" sounds ambiguous in a similar way). The notion that it would turn out that water is XYZ if the XYZ-world turned out to be actual is entirely compatible with the claim that water is necessarily H₂O. Again, this is simply a commonsense reading of the "turns out" locution, a reading on which it behaves differently from a standard modal locution.]]

*[[One must be careful to ask the question from the *personal* point of view, rather than from a community's point of view, to handle the cases of semantic deference discussed below. That is, we must adopt the (reasonable) reading of the locution on which I can say truly (in a case where I defer to my community) that if it turned out that the community uses "elm" to refer to X-trees, it would turn out that elms are X-trees.]]

There is often an element of indexicality in the fixation of reference, so the location of a center must be specified for reference in an actual-world candidate to be determined. For example, if it is specified that an actual-world candidate contains both H₂O and XYZ on otherwise similar planets, the referent of my concept "water" is undetermined until we specify where in that world my "viewpoint" - the center - is located. This indexicality is most obvious for a concept such as "I", whose primary intension picks out the individual at the center of a centered world. In general, an actual-world candidate is a world-as-I-find-it, marked with a viewpoint at its center.[*]

*[[Actual-world candidates resemble the "contexts of utterance" of Kaplan 1989, although the resemblance is imperfect. Counterfactual-world candidates, discussed below, correspond to Kaplan's "circumstances of evaluation".]]

The primary intension of a concept is determined independently of the concept's actual referent; it must be, as the intension itself specifies the way in which the referent depends on how the world turns out. This is reflected in the *a priori* thought-experimental methodology in the investigation of primary intensions. Given my concept "Gödel", I can ask (with Kripke): to what would it refer if it turned out that a man named "Schmidt" had proved the incompleteness of arithmetic, before the proof was stolen and published by a man named "Gödel"? Empirical knowledge about the actual world is irrelevant here, as specification of the basic facts about the centered world in question gives me all the facts that could possibly be relevant.

It is a central fact about concepts that for a wide range of actual-world candidates, the referent of a concept in such a world is both determinate and accessible to the thinker. To see this, note that once we know all about the qualitative structure of the actual world and our location within it, we have all the information we need to ascertain the nature and existence of our concepts' referents (think of "Hesperus", "water", "Jack the Ripper"). And if the actual world had turned out differently, we could have done the same. Information about the world's structure and our location within it would have sufficed, in principle, for knowledge of reference. Given this ability, we can do the same in cases of counterfactual speculation about actual-world candidates, provided that we are supplied with the relevant information about those worlds. Indeed, such thought-experiments are the basic currency of the theory of reference. The primary intension of a concept is therefore accessible in principle to the thinker, and accessible prior to investigation of the thinker's external world.

We need not require that an actual-world candidate itself contains a copy of the concept in question. A primary intension specifies what it takes for an entity in the actual world to qualify as the referent of the concept; these conditions of application will often build in no requirements about the presence of the concept itself. In evaluating the referent at an actual-world candidate, we retain the concept from the *real* actual world. For example, in an actual-world candidate in which the individual at the center is in a coma, the primary intension of my concept "I" (from the real actual world) picks out the comatose individual.

Intuitively, we would say that if this world turned out to be actual, it would turn out that I am comatose, in the same sense in which if the XYZ-world turned out to be actual, it would turn out that water is XYZ. Not all concepts are so straightforward. We will see later that the primary intensions of some concepts are not determinate across all centered worlds, picking out a referent only if certain background conditions are satisfied. All that matters for now, however, is that the primary intension of a concept is a non-trivial partial function, determinate across a large class of worlds.

The *secondary intension* of a concept is more familiar, but works quite differently. This is a function $f_2: W \rightarrow R$ from uncentered worlds to referents, returning the referent of the concept in a world when it is *considered as counterfactual*, given that the actual world has already been fixed. To evaluate the secondary intension of one's concept "water" at a world w , one asks: what is water in w ? The same goes for other concepts, *mutatis mutandis*.[*] In the case of "water", if we accept the intuitions of Kripke and Putnam, then given that water is H₂O in the actual world, water is H₂O in all worlds, so the secondary intension of the concept picks out H₂O in all worlds. Similarly, if Putnam and Kripke are correct, I am DJC in all counterfactual worlds, so the secondary intension of my concept "I" picks out DJC in all worlds. Note that these counterfactual worlds require no center; in evaluating the secondary intension at a world, the viewpoint of a thinker within that world is irrelevant.

*[[So the rules-of-thumb for determining primary and secondary intensions each have two formulations, depending on whether the concept is mentioned or used. On the "mention" formulation: the primary intension picks out the referent of a concept in w when it is considered as actual, whereas the secondary intension picks out the referent of a concept in w when w is considered as counterfactual. On the "use" formulation, the primary intension picks out what (for example) it would turn out that water is if w turned out to be actual (where the "turns out" locution is interpreted appropriately), whereas the secondary intension picks out what water is in w .]]

Unlike the primary intension, the secondary intension is often determined by *a posteriori* factors, as it depends on how the world of the thinker turns out. Given that the liquid in the oceans and the lakes in this world is H₂O, the secondary intension of my concept "water" picks out H₂O in all worlds, but if the liquid had turned out to be XYZ, then the secondary intension would have picked out XYZ in all worlds instead.

The full story about a concept can be summed up in a *two-dimensional intension* $F: W^* \times W \rightarrow R$, where $F(v,w)$ is the referent of the concept in w , if v turns out to be actual (for $v \in W^*$ and $w \in W$). The primary intension is $f_1: v \mapsto F(v,v)$, where $v \in W^*$ and where v' is the uncentered world obtained by "unmarking" the center of v . The secondary intension is $f_2: w \mapsto F(a,w)$, where a is the actual world of the thinker. Because f_2 depends on a , it is in general accessible only *a posteriori*. Note that the primary intension and the secondary intension always pick out the same referent in the actual world: $f_1(a)=f_2(a)=F(a,a)$.

The secondary intension of a concept is usually determined by the primary intension and the actual world, but the way in which it is determined depends on the concept involved. There are a few basic patterns of determination, corresponding to different types of concepts.[*] The simplest case is that of ordinary "descriptive" concepts - perhaps "doctor" and "circle", although semantic intuitions differ - for which the

conditions of application in a counterfactual world are the same as they are in the actual world. For these concepts, the secondary intension is simply a copy of the primary intension, over the space of uncentered worlds: $f_2(w) = f_1(v)$, where $v' = w$. This requires $f_1(u) = f_1(v)$ for all $u, v \in W^*$ such that $u' = v'$ - that is, the value of the primary intension at a centered world cannot depend on the position of the center - so the primary intensions of such concepts cannot be in any way indexical. In these cases, the secondary intension is independent of *a posteriori* factors.

*[[I count at least four basic patterns: those for descriptive concepts, rigid designators, natural kind concepts, and semantically deferential concepts. There may be others. The pattern itself is independent of *a posteriori* factors; for example, the fact that we use "water" to pick out in counterfactual worlds whatever has the common underlying structure of the stuff in the actual world (if such a common structure exists) is conceptually prior to investigation of the empirical details. We can think of these patterns as written into the *a priori* two-dimensional structure of a concept.]]

More interesting are cases in which the secondary intension differs significantly from the primary intension, and is determined only *a posteriori*. The most straightforward version of this phenomenon occurs with concepts corresponding to rigid designators, such as indexicals and names. For such concepts, the secondary intension is determined by *rigidifying* the primary intension - that is, by evaluating it at the actual world, and picking out the same referent in all possible worlds. For example, it turns out that the primary intension of my concept "I" picks out DJC at the actual world, so the secondary intension picks out DJC at all worlds in which he exists. A similar dependence, with minor complications, holds for natural kind terms.[*]

*[[Strictly speaking, natural-kind concepts are not rigid designators: "cat" does not pick out the same referent at every world, as there are different cats in each. One way around this is to stipulate that "cat" picks out the property of cathood at every world. Less artificially, we can give a two-stage story according to which "cathood" picks out the same property in each world, and "cat" picks out the objects satisfying the property in that world. Or we might simply say that "cat" picks out the objects with the same underlying explanatory structure as the cats in the actual world. I will ignore these complications in what follows.]]

For rigid-designator concepts, the structure of the secondary intension is easy to characterize, but the structure of a primary intension is generally more complex. Often the best one can do is give a rough summary. For example, it is not too far off the mark to say that the primary intension of my concept "water" picks out the dominant clear, drinkable liquid in the environment that stands in an appropriate causal relation to the center of a world, but this is far from perfect. One can characterize the primary intension more carefully by considering difficult cases (what would the referent be if there was both XYZ and H₂O in the lakes? what would it be if the liquid that falls from the sky turns out to differ from the liquid we drink, due to a complex reaction? what if most "liquids" turned out to be made of fine particles of sand?), but we cannot expect to arrive at a perfect characterization of any manageable length. Indeed, a perfect finite characterization may be impossible. It is sometimes useful to appeal to rough-and-ready characterizations for the purposes of illustration, but it must be stressed that the real object of interest is always the full intension, to which any such descriptive characterization is a mere approximation.

It should be noted that this picture is entirely neutral between "causal" and "descriptive" theories of reference. On *any* account of how reference is fixed, there will be some way in which the referent of a

concept depends on how the actual world turns out: if the actual world turns out this way, then "elm" refers to these things; if it turns out another way, then "elm" refers to those things. That is, on *any* theory of reference a concept will have a primary intension. The subject matter of the theory of reference is the precise structure of primary intensions; this is something about which the present account is neutral.[*] When I occasionally characterize a primary intension in descriptive terms, this is only an approximation for the purposes of illustration, and nothing will depend on the details.

*[[Some may think that an account involving primary intensions "feels" like a description theory, and perhaps it does. But note that even Kripke's (1980) argument against description theories *presupposes* something like a primary intension. It proceeds by considering the referent of a concept such as "Gödel" at various actual-world candidates, thus evaluating the primary intension, and arguing that any given description gives the wrong result. It follows that either an account involving primary intensions need not be a description theory, or there is a class of description theories that Kripke's arguments do not refute. The choice between these conclusions presumably depends on just what one takes a description theory to be.]]

The framework applies equally to concepts that are used with *semantic deference*. As Putnam and Burge have pointed out, my concepts of "elm" and "arthritis" can refer appropriately despite my knowing very little about the relevant subject matter. In these cases, I refer via the division of linguistic labor. At a first approximation, if those around me (and especially the "experts" in my community) refer to X-trees by "elm", then my concept "elm" refers to X-trees; if they refer to Y-trees by "elm", then I refer to Y-trees, and so on. The concept therefore determines a perfectly good primary intension. To characterize this primary intension even approximately, we have to involve the *name* of the concept: thus, at a first approximation, my concept "elm" picks out whatever trees those around me (particularly experts) refer to as "elms". As always, reference of a concept depends on how the actual world turns out; it is just that in these cases, some of the relevant facts about the actual world are linguistic.

Putnam suggests that concepts like "water" and "elm" show that the intension of a concept cannot determine the extension, if an intension is internally determined. The two-dimensional analysis shows that this is only half-true. The primary intension of a concept determines its extension, and the primary intension is internally determined. Of course, the primary intension is a *centered intension*, taking a centered world as argument, so Putnam's claim still holds for uncentered intensions. But any intension requires facts about the actual world to determine extension, and it is most natural to regard the actual world of a thinker as centered, so it is no surprise that centered intensions should be central in the fixation of reference.

In some ways, the primary intension of a concept resembles a Fregean sense. It is the primary intension that fixes reference in the actual world, and the primary intension is generally cognitively accessible. We will see that instances of Frege's puzzle arise precisely when the concepts involved have different primary intensions, so insofar as senses are introduced to solve these problems, primary intensions can play the role. On the other hand, Frege required that senses determine reference non-indexically, stipulating that the sense of a concept such as "here" is tied to a particular place. Here, the secondary intension seems to be a closer fit. Many have noted that nothing can satisfy all Frege's requirements on sense,[*] and it may be fruitful to view the present account as dividing the notion of sense into two distinct components. Alternatively, one might make the case that Frege's requirement of non-indexicality is misguided, and that

primary intensions can do much of the important work that senses are introduced to do, despite this and other differences of detail with Frege's account.

*[[E.g., Perry 1977.]]

There should be no question of whether the primary or secondary intension is *the* intension associated with a given concept. The full story can only be given two-dimensionally. One or the other may be more useful for various specific purposes. In matters of linguistic content, the secondary intension may play the major role, as different users of a term may have quite different associated primary intensions, whereas the secondary intension will generally be constant. This is particularly clear in the case of names, for which primary intensions can vary wildly across a community. For questions about thought and its role in governing action, however, we will see that the primary intension is central.

3 Notional and Relational Content

So far we have been concerned with the semantics of concepts. Now we move to the central topic, the semantics of thoughts, which stand to concepts roughly as sentences stand to terms. As before, I will be concerned with thoughts (usually beliefs) that are associated with sentences in the thinker's language, but the account generalizes. Just as concepts are associated with intensions, thoughts are associated with *propositions*, which here will be regarded as functions from possible worlds into truth-values, or more simply as sets of possible worlds. (The framework can be adapted to other accounts of propositions, e.g. on which they are seen as structured groups of intensions.) Propositions be regarded as the contents of thoughts and of sentences, just as intensions can be regarded as the contents of concepts and of terms.

Just as there are two intensions associated with every concept, there are two propositions associated with every thought. The *primary proposition* is a *centered proposition* - a set of centered worlds, or a function from centered worlds to truth-value. At a centered world w , it returns the truth-value of the thought in w when w is considered as actual.[*] As a heuristic, we can ask: would the thought be true, if w turned out to be actual?[*]

[[Stalnaker (1978) introduces a proposition like this as the "diagonal proposition" associated with an utterance. Just as the primary intension can be seen as the "diagonal" of a two-dimensional intension - $f_1(v) = F(v, v')$ - the primary proposition can be seen as the diagonal of a "two-dimensional proposition", or a function from $W^ \times W$ into truth-values. For reasons that will become clear, however, I find this terminology misleading in its suggestion that primary intensions and propositions are in some way derivative on secondary intensions and propositions.]]

*[[As before, we might also ask questions analogous to "if w turned out to be actual, would it turn out that water is wet?", where the "turns out" locution is read as before. As before, with either heuristic we retain the thought from the real actual world.]]

For example, the primary proposition associated with my thought "I am a philosopher" is the set of centered worlds in which the individual at the center is a philosopher. Or consider my thought "There is water on this table". If the actual world turned out to contain XYZ as the dominant liquid in the oceans

and lakes, and if there were a table in front of me supporting a glass of XYZ, then my thought would be true. As things really are, the world contains H₂O in the oceans and lakes, and there is only an empty glass on the table in front of me, so my thought is false. At a rough approximation, the primary proposition associated with my thought is the set of centered worlds in which there is a sample of the dominant clear liquid in the environment on a table in front of the individual at the center, although as usual the approximation is no substitute for the proposition itself, which can be obtained only from careful consideration of cases.

The *secondary proposition* is an uncentered proposition, returning at w the truth-value of the thought at w when w is considered as counterfactual. As a heuristic, we can ask questions of the form: "in w , is it the case that water is wet?", and similarly *mutatis mutandis* for other thoughts. Like secondary intensions, secondary propositions are often determined only *a posteriori*, as they depends on how the actual world turns out. For example, the secondary proposition of my thought "I am a philosopher" is the set of worlds in which DJC is a philosopher. Or take my thought "there is water on this table". Given that the actual world turns out as it has, with H₂O in the oceans, my study table T in front of me, and with my thought occurring at time t , the secondary proposition is that set of worlds in which there is a sample of H₂O on table T at time t .

The primary proposition associated with a thought can usually be determined by composing the primary intensions of the concepts involved.[*] For example, the primary proposition of a thought "Hesperus is Phosphorus" will be true at a centered world if the individual picked out by the primary intension of "Hesperus" in that world is also picked out by the primary intension of "Phosphorus" in that world - that is, roughly at those worlds in which the morning star is the evening star. A similar relation holds between secondary intensions and secondary propositions. The secondary intensions of "Hesperus" and "Phosphorus" pick out the same object in all worlds in which the referent of either exists, so the secondary proposition of "Hesperus is Phosphorus" is true of all such worlds.

*[[This does not require a "language of thought": it requires only that thoughts are semantically composed of concepts, not that they are syntactically composed of concepts. Even the assumption of semantic compositionality is inessential, although it eases discussion. The definition of primary and secondary propositions requires only that thoughts are the kind of thing that can be true or false of the world.]]

I will call the primary proposition associated with a thought the thought's *notional content*, because it most directly reflects how things seem from the point of view of the thinker. I will call the secondary proposition the thought's *relational content*, as it often (although not always) depends on the relations that the thinker bears to objects in the external world. For example, when you think "I am hungry" and I think "You are hungry", our thoughts have different notional content, but similar relational content. When we both think "I am hungry", the situation is reversed.

Importantly, the notional content of a thought is determined entirely by the internal state of the thinker.[*] Like the primary intension, the primary proposition cannot depend constitutively on the external world, as the proposition itself specifies how the truth of a thought depends on how the actual world turns out. I can investigate the primary proposition associated with my thought without investigating the environment: all

I need to do is ask, *if* the environment turns out like so, will my thought be true? This is an entirely internal enterprise; facts about the actual environment have no role to play. Of course, the notional content of a thought will almost always depend *causally* on the external world, but whenever the external world affects the notional contents of our thoughts, it will do so by affecting the internal state of the thinker.

*[[Thus even a brain in a vat might have thoughts with notional content. This can be used to defuse Putnam's (1981) anti-skeptical argument that if he were a brain in a vat, he could not think "I am a brain in a vat". A brain in a vat could think a thought with the appropriate notional content, if not the appropriate relational content, and the notional content of the thought would seem sufficient to express a significant skeptical possibility, true only in worlds in which the individual at the center lacks the usual sort of epistemic contact with the surrounding world.]]

Every thought has both notional and relational content. As notional content is internally determined, we may stipulate that a thought's *narrow content* is its notional content. Similarly, we can stipulate that a thought's *wide content* is its relational content. It should be noted, however, that relational content is not always dependent on the environment. In particular, if a thought involves only non-rigid concepts used without semantic deference, the associated secondary proposition will be a simple twist on the primary proposition, so that relational content will be internally determined.[*]

*[[For similar reasons, this differs from Quine's usage on which "notional" and "relational" belief properties correspond to those attributed by *de dicto* and *de re* ascriptions respectively. *De re* belief properties and relational content come apart where nonrigid concepts are concerned: the relational content of my belief that the shortest spy is a spy might be shared by many twins in different environments, but the corresponding *de re* belief property requires a specific referent. Further, *de dicto* belief ascriptions often ascribe a combination of notional and relational content, as I discuss later.]]

As promised, this sort of narrow content is truth-conditional. The notional content of a thought delivers the conditions a centered world must satisfy, in order for the thought to be true in that world. The notional content of my thought "There is water on the moon" is true in a centered world with XYZ in the oceans and lakes and XYZ on the moon, even though the relational content of my thought is false there. At a rough approximation, the notional content will be true in those centered worlds in which the dominant clear liquid found in the oceans and lakes surrounding the center can also be found on the heavenly body bearing the appropriate relation to the center. We can think of the notional truth-conditions of a thought as expressing the way the actual world has to be, in order for the thought to be true.

The notional truth-conditions of a thought are often very different from the relational truth-conditions. One might worry that a thought could then turn out to be both true and false, but this is impossible. Just as the primary and secondary intensions pick out the same referent in the actual world, the primary and secondary propositions share the same truth-value there, so there is no danger of a thought's being both true and false in actuality. It is only at other worlds that the truth-conditions will give different results, and this is a consequence of the way those worlds are treated. In evaluating the primary proposition they are considered as actual, whereas in evaluating the secondary proposition they are considered as counterfactual. We might say that notional truth-conditions tell us what it is for a thought to be true *in* a world, whereas relational truth-conditions tell us what it is for a thought to be true *of* a world.

The distinction between notional and relational truth-conditions parallels Kripke's distinction between the

a priori and the necessary. A thought is knowable *a priori* if its notional content is true in all centered worlds; and it is necessarily true, in Kripke's sense, if its relational content is true of all worlds. Cases of the contingent *a priori*, such as the thought that the stick in Paris is a meter long if it exists, are cases in which a thought has a necessary primary proposition but a contingent secondary proposition. Cases of the necessary *a posteriori* are the reverse.[*]

*[[A related two-dimensional account can be given of the corresponding phenomenon in language, yielding in a certain sense a semantic explanation of *a posteriori* necessity. I develop such an account in Chalmers 1996, where it is used to analyze some applications of *a posteriori* necessity in the mind-body problem. Upon analysis, the two-dimensional account suggests that some of the more ambitious applications of *a posteriori* necessity (in defence of certain sorts of materialism and of certain sorts of moral realism, for example) are unlikely to succeed.]]

4 The Advantages of Notional Content.

In recent times, the "content" of a thought has generally been identified with its relational content; but the notional content seems to be an equally good candidate. As before, there is no need to decide which is *the* content; but that being said, there many ways in which the notional content of a thought is responsible for most of the explanatory work that we would expect a notion of content to do.

First, notional content determines the rational relations between thoughts. If one thought implies another thought *a priori*, it is because the primary proposition associated with the first implies the primary proposition associated with the second. If I know that it is hot where I am now, I know that it is hot here, and vice versa; this is reflected in the fact that the notional contents of the two thoughts are the same. The relational contents of the thoughts are very different, however: there is no obvious relation between the proposition that it is hot where DJC is at time *t* and the proposition that it is hot at place *P* then.

It is straightforward to see why this is so. If one thought implies another *a priori*, then in any centered world in which the first turns out to be true, the second will turn out to be true, so the notional content of the first will imply the notional content of the second. Conversely, if the notional content of one thought implies the notional content of another, then a thinker should in principle be able to infer one from the other *a priori* (setting aside problems with complex mathematical statements and the like[*]), as the relationship between the thoughts is independent of facts about the external world. This is not so for relational content: the relational content of a thought may depend on facts about the external world that are unknown to the thinker, so that relations between thoughts in virtue of their relational content may lie outside the realm of cognitive access.

*[[The hard cases involve thoughts that are true in all contexts, but whose truth is extremely difficult for us to determine; perhaps Goldbach's conjecture is such a case. How these cases should be treated depends on whether we regard our limitations here as arising from imperfect rationality. If so, the link between notional content and rational relations remains precise; if not, it needs a small qualification. In any case, this issue arises only in rare cases.]]

This can be applied straightforwardly to explain the informativeness of a statement such as "Hesperus is Phosphorus". Although its associated secondary proposition is equivalent to that of the trivial "Hesperus

is Hesperus", its associated primary proposition is quite distinct, so it is not cognitively trivial. If a thought is notionally equivalent to a thought that is knowable *a priori*, it will itself be knowable *a priori*,[*] but an equivalence in relational content carries no such guarantee.

*[[Subject to the qualification in the previous note.]]

We can also invoke notional content in the case of Kripke's Pierre, who paradoxically seems to believe that London is pretty and that London is not pretty, without any breakdown in rationality. The two-dimensional framework shows why there is no paradox. Pierre's concepts "Londres" and "London" have quite different primary intensions - in a given centered world, the first picks out (approximately) the famous city called "Londres" that the individual at the center has heard about, whereas the second picks out the grimy city in which that individual has been living. The secondary intensions are identical, of course, picking out London in every world. Pierre's two beliefs, then, have contradictory associated secondary propositions, but their primary propositions are compatible. Contradictory secondary propositions support no charge of irrationality, as rational relations are determined by the first dimension of content.

Intuitively, Pierre's two beliefs are rationally compatible because there is some way the world could turn out such that both are true. There is a centered world in which "Londres" names a faraway, beautiful city (maybe it is in India), and in which the individual at the center inhabits an entirely distinct ugly city called "London"; and for all Pierre knows and believes, such a world could be actual. As long as there is such a world, satisfying the notional content of all Pierre's thoughts - that is, as long as the *notional* contents of his thoughts are compatible - his rationality is not in danger.

This brings out the relation between this account and Dennett's (1981) suggestion that the narrow content of a thought is reflected in the *notional world* of the thinker. The notional world we can take to be a world (really a class of worlds) in which all of a thinker's beliefs (or as many as possible, in cases of irrationality) would be true.[*] A notional world is to be thought of as a world as the thinker finds it, so of course it is a centered world; and beliefs are to be evaluated in that world according to their primary propositions. Pierre's notional world is a world in which there is a beautiful faraway city called "Londres", and a grimy city close at hand called "London". If Pierre really lived in his notional world, he would be right about everything and rarely surprised.

*[[Dennett suggests that the relevant worlds are "the environment (or class of environments) to which the organism as currently constituted is best fitted." I think that this gives imperfect results in some cases, most obviously for beliefs about the past; see also the criticisms in White 1991, and the more refined account put forward there. Dennett's and White's suggestions here are in effect first attempts at giving a naturalistic reduction of notional content. Such a reduction is likely to be a major project in its own right.]]

On similar grounds, one can make the case that notional content governs the *cognitive* relations between thoughts. Here there is an important qualification, however, as notional content as I have defined it does not distinguish the various cognitive relations that might hold between thoughts that are deductively equivalent. From the point of view of notional content, a complex mathematical proof is as trivial as

modus ponens; so the fine-grained cognitive dynamics of deduction lies beyond the reach of notional content. Perhaps a refined variety of notional content might handle these cases, but in any case I will set these issues aside here, as relational content does not handle them any better, and they are largely independent of the issues at play in this paper.

A qualified thesis would be the following: insofar as notional content or relational content governs the cognitive relations between thoughts, it is notional content that does the work. This can straightforwardly be seen by examining cases in which thoughts are cognitively related, and by observing that (1) in related cases in which relational content of these thoughts is varied but notional content is held constant, the cognitive relations are preserved (except insofar as cognitive relations can be affected by varying factors independent of *both* notional and relational content, as in the deductive case), and that (2) in cases in which the relational content is preserved but the notional content is not, the cognitive relations are often destroyed. The details here parallel those of the discussion of the explanation of behavior, below, so I will not duplicate them here.

A third advantage of notional content is its suitability for a role in the explanation of behavior. It is often noted that relational content seems slightly out of synchrony with what one would expect of an explanatory psychological state. To use an example of Kaplan's, [*] if you are watching me and my pants catch fire, our respective beliefs that my pants are on fire now will have the same relational content (they are true in all worlds in which DJC's pants are on fire at time *t*), but will lead to very different actions (I might jump into a river, while you just sit there). The difference between our actions does not seem to be something that a characterization in terms of relational content alone can explain. In a similar way, belief states can produce very similar behavior for apparently systematic reasons, even when the beliefs have very different relational content - witness the behavior that my twin and I produce when we think about twin water and water respectively, or the similarity between the actions of two people who think "I am hungry". A whole dimension of the explanation of behavior seems out of the reach of relational content.

These explanations can be easily handled in terms of notional content. If you and I think "I am hungry", the notional contents of our thoughts are very similar - our thoughts are associated with similar primary propositions, although different secondary propositions - and that similarity is reflected in the similarity of our actions. When you and I both believe that my pants are on fire, on the other hand, our notional contents are very different, and our actions differ correspondingly. Note that this provides a straightforward solution to Perry's problem of the essential indexical: it is notional content, not relational content, that governs action, and notional content, unlike relational content, is a centered proposition. [*]

*[[Perry (1979) considers the possibility that centered ("relativized") propositions might provide a solution, but dismisses it on the grounds that believing that such a proposition *P* is "true for me" does not distinguish me from third parties who also believe that *P* is true for me, but act differently. The trouble is that Perry's locution "true for me" introduces an unnecessary element of relational content. What distinguishes me from the third parties is rather that I believe *P simpliciter*, or better, that my belief has *P* as its notional content.]]

Notional content also accounts for the similarity of action between twin cases; this similarity reflects the fact that my beliefs about water and my twin's beliefs about twin water have the same associated primary

propositions. But we need not move to the realm of science fiction to see the point. Two thoughts can share notional content even when two thinkers are quite different, as our thoughts "I am hungry" show, and even in these cases, similarities in narrow content will lead to similarities in action, other things being equal. Say I think that Superman is across the road, and I want to have Superman's autograph; then other things being equal, I will cross the road.[*] If you have thoughts with similar notional content to mine, then you will do the same. But if your thoughts only share relational content with mine, but have different notional content - say you think that Clark Kent is across the road, but want Superman's autograph -- then your behavior may be correspondingly quite different.

*[[To simplify the discussion, I make the happy assumption that Superman is actual and is identical to Clark Kent.]]

In general, whenever the content of a thought is causally relevant to behavior, its contribution is screened off by that of notional content in the following sense: if the thought had had the same notional content but different relational content, the behavior would have indiscernible (except insofar it might be affected by changing factors independent of both sorts of content), whereas if it had had the same relational content but different notional content, the behavior might have been quite different.

To see the latter point, we need only examine cases like those above. The thoughts "I am hungry" and "The guy over there is hungry" (unknowingly looking in a mirror) will lead to very different behavior, even though their relational content is the same. When Lois Lane is trying to cut Clark Kent's hair, her observation "Clark's hair breaks the scissors" will prompt a reaction very different from that provoked by a corresponding thought concerning Superman. If I hear that Cary Grant is starring in a movie, I might be more likely to watch than if I hear that the movie stars Archie Leach. In all these cases, different reactions are provoked by a difference in the notional content of a thought. In general, whenever the notional content of a thought is varied, different consequences can be expected, even if relational content is preserved throughout.[*] Given that notional content governs cognition and that cognition governs action, this is just what we would expect.

*[[Of course, thoughts like "Cary Grant is in the movie" and "Archie Leach is in the movie" might lead to the same actions despite their different notional content, if I know that Cary Grant is Archie Leach. But even here, there exist circumstances under which the thoughts might play a different role - if someone tells me that Cary Grant is not Archie Leach after all, for instance. In general, whenever two thoughts have different notional content, there are at least hypothetical circumstances under which the action-governing roles of the thoughts will differ.]]

By contrast, if the relational content of a thought is varied but notional content is kept constant, behavior stays indistinguishable throughout. Perhaps, unbeknownst to me, Cary Grant is an elaborate hoax, a co-operative construction by avant-garde animators and the mass media. In such a case, my thought about Cary Grant will have no relational content at all, but as long as it has the same notional content my behavior will be indistinguishable from that in the case in which he is real. Or perhaps Cary Grant is really Ludwig Wittgenstein in disguise: if so, very different relational content, but the same behavior. Similarly, when my twin and I think "I need some more water for this pot", the relational contents of our thoughts differ, but we both rush for the sink.

We can make a similar point within a single system. To use an example of Evans', let Julius be a name that functions to rigidly designate whoever invented the zip. Then the primary intensions of my concepts "Julius" and "the inventor of the zip" will be the same, but the secondary intensions will be very different. Despite the difference in secondary intensions, however, it is clear that any thoughts that "Julius is *P*" will play the same role in directing cognition and action as thoughts that "The inventor of the zip is *P*". The rigidification and consequent difference in secondary intension is simply irrelevant.[*]

*[[With one slight exception: the two concepts will behave differently in modal thought, as when one judges that it is necessary that Julius is Julius but not necessary that Julius is the inventor of the zip. But even here the difference is accounted for by a difference in the internally determined two-dimensional intension, rather than by a difference in relational content *per se*.]]

Some might object that there are cases in which we individuate behavior extrinsically - for example, my twin and I might behave differently in that I drink water where he drinks twin water - so that there is a dimension of behavior that escapes notional content. But even in this sort of case, relational content does not usually help. Even my twin, with the different relational content, would drink water if he was in my present context. What is relevant to behavior here is not relational content but current context, as we can see by an extension of the varying-factors strategy; and I certainly do not wish to deny that current context is relevant in the explanation of behavior.

The only cases in which there is a direct tie between relational content and behavior are cases in which behavior is individuated by an intentional object, such as that in which we say that *search for* a glass of water whereas my twin searches for a glass of twin water. This connection holds across all contexts, as behavior only counts as water-searching if it is caused by water-thoughts. But for the same reason, this is a very weak sort of relevance for relational content: as Fodor (1990) notes, the relational contents of thoughts are not *causally* relevant to action here, but instead are *conceptually* relevant, in effect determining the category the action falls under.[*] And relational content gives us very little purchase in the *explanation* of action here, as we will only know that some behavior is water-searching if we already know that water-thoughts lie behind it. In a causal (as opposed to a conceptual) explanation of the action, notional content will still play the central role.

*[[See Fodor 1990 for a detailed argument along these lines. I note also that one can individuate this sort behavior intentionally but still narrowly if one individuates intentional objects notionally.]]

Why is notional content primary? To answer this question, it is useful to think of my belief contents as constituting a model of my world, a kind of map by which I steer. My model is a model of the world as I find it, a centered world with me at the center, and my beliefs are constraints on that world. One belief might rule out *these* centered worlds as a candidate for the world where I am, another might rule out *these*, until only a limited class of worlds is left. I operate under the assumption that my world is one of those worlds, and if I am lucky I will not be too surprised.

My world-model is ultimately a *notional* world: a set of centered worlds, such that none of these would overly surprise me if they turned out to be actual. The constraints on this model are those of notional

content. Relational content puts some additional constraints on the world, but these are not constraints that are useful to me. The relational content of my belief that the liquid in thermometers is mercury endorses only those worlds in which thermometers contain the element with atomic number *X*, but this constraint is so deeply inaccessible to me that if it turned out that the liquid has atomic number *Y*, I would not be in the least surprised. In an important sense, this constraint is not reflected in my world-model at all. Insofar as my world-model is useful to me, the constraints on it are entirely those of notional content; and insofar as content is useful to me, it is notional content that carries the burden.

It is worth noting that in making a case for the primacy of notional content, I have not appealed to any *a priori* methodological principles such as the dictum that what governs behavior is in the head. The case for notional content has been made directly, independently of questions about physical realization. Indeed, it should be stressed that nothing I have said implies that facts about a thinker's environment are irrelevant to the explanation of behavior. Facts about the proximal environment will clearly play an important role insofar as they affect the thinker;[*] facts about the current environment are crucial to explaining the success or failure of various actions; and facts about environmental history will at least be central to a causal explanation of a thinker's current cognitive state. All that follows from the present framework is that the environment is not relevant to the explanation of behavior *in virtue of its role in constituting relational content*. The kind of *content* that governs behavior is purely notional.

*[[It may even be that in certain cases, notional content can be constituted by an organism's proximal environment. For example, one could argue that the contents of a notebook with a list of addresses can constrain my notional world in the same way that my memory might. In these cases, one is effectively individuating the cognitive system as a coupled system whose boundaries lies beyond the skin, with notional content internal to the *coupled* system. Of course one could always keep the notional contents within the skin by insisting that the relevant contents are of the form "the address in this notebook" rather than "53rd Street", but in some cases one may gain explanatory purchase by taking the broader view of the system and the more constrained view of its notional worlds. This kind of consideration leads to an "externalism" very different from that deriving from considerations about relational content, and arguably more relevant to the explanation of behavior; it is discussed in much more detail by Clark and Chalmers (1995). Of course notional content remains internal in the deepest sense; it is just that the skin is not a God-given boundary of a cognitive system. Still, this is another way in which the issue between notional and relational content runs deeper than the issue between internalism and externalism.]]

5 Belief Ascription and Psychological Explanation

This raises something of a puzzle about the role of belief ascriptions in psychological explanation. If what has gone before is correct, the kind of content that governs cognition and action is notional, and therefore narrow. But at the same time, there is strong evidence that the kind of content attributed by belief ascriptions is at least sometimes wide. Does this mean that the common-sense framework of explanation in terms of belief ascription should be discarded? Alternatively, is the apparent success of the common-sense framework evidence that something in these arguments has gone badly wrong?

Neither conclusion is justified. The present framework shows how it can at once be true that (1) belief ascriptions ascribe wide content, (2) narrow content governs action, and (3) belief ascriptions explain action. In short: Belief ascriptions ascribe a combination of notional and relational content. It is in virtue of the relational component that the ascribed content is wide, and it is in virtue of the notional component

that the ascribed content is explanatory.

A full justification of this answer requires two things: first, an analysis of what is attributed in belief ascriptions, so that we can see precisely what sorts of notional and relational content are attributed; second, an analysis of the role of belief ascriptions in psychological explanation, so that we can see that even in ordinary practice, it is the notional content attributed that carries the explanatory burden. I cannot provide anything like a complete treatment of these matters - the analysis of belief ascriptions deserves entire volumes of its own - but I can provide a preliminary sketch.

It is easy to see that ordinary belief ascriptions ascribe both notional and relational content. If I say "Ralph believes that Clark Kent is muscular", in order for my utterance to be true Ralph must have a belief that satisfies two sorts of constraints. First, the belief must have the relational content of the proposition that Clark Kent is muscular (perhaps we can allow a certain amount of variation in the relational content, if for example his concept of muscularity is slightly different from the norm). But that alone is not enough: a belief that Superman is muscular would have the same relational content, but would not make my ascription true. For the ascription to be true, his belief must involve a concept that refers to Clark Kent under an appropriate primary intension: perhaps an intension that picks out whoever is called "Clark Kent", or one that picks out whoever is that reporter with glasses at the Daily Planet, or more likely some complex intension in the vicinity. A variety of primary intensions will do, in order for the ascription to be true, but still a narrow class compared to all the primary intensions that in fact refer to Clark Kent.[*] A rigidified concept whose primary intension picks out the guy in the cape, or the strongest man in the world, might refer to Clark Kent, but the thought that the person in question is muscular would not make my ascription true.

More straightforwardly, if I am right in saying "Tom believes that he is hungry", then Tom must have a belief with more or less the appropriate relational content, true of all those worlds in which Tom is hungry at time *t*, but there is a strong constraint on notional content too. In particular, Tom must refer to himself via the primary intension that picks out the individual at the center in every centered world. If he sees someone in the distance clutching their belly, without realizing that he is in fact looking into mirrors, then a thought that that person is hungry has the right relational content, but on the most natural reading it does not make my ascription true. The ascription will only be true if Tom's belief refers to himself *as* himself; that is, only if its notional content is appropriate.

We have seen that content decomposes naturally into notional and relational content; we now see that belief ascription puts strong constraints on both. The precise nature of these constraints is the subject matter of a theory of belief ascription. Ideally, such a theory will yield, for any given ascription, the conditions that a beliefs' notional and relational contents must satisfy in order to make the ascription true. We can think of a belief ascription as marking out a subspace in the space of (notional content, relational content) ordered pairs.

The constraints on relational content are generally fairly straightforward, as the examples above show. Usually, these constraints can be derived from the reference or meaning of the terms involved in the ascription. The constraints on notional content are vaguer and more subtle. Certainly not every primary

intension referring to Clark Kent qualifies a concept to play a role in making a "Clark Kent" belief ascription true, but the precise restrictions on what counts as a "Clark Kent" primary intension are not at all obvious. Similarly, to believe that water contains hydrogen atoms, it presumably does not *suffice* to have a relevant belief about H₂O (imagine a very theoretical chemist who *only* knows about H₂O's chemical properties); a "water"-appropriate primary intension is required, but just what are the conditions here? This is a matter for careful linguistic investigation. The constraints on notional content will be somewhat indeterminate in intermediate cases, and they will probably be context-sensitive, but they will be there, and they will generally be quite strong.

This provides a solution to the fifth puzzle mentioned at the start. The "modes of presentation" central to a theory of beliefs ascription can be seen as primary intensions and primary propositions.[*] To satisfy a belief ascription, a believer must not only have a belief with the appropriate (secondary) proposition, but must also believe it under an appropriate mode of presentation in this sense.[*] As a bonus, the fact that modes of presentation are themselves intensional or propositional allows for a certain elegance and symmetry in the theory of belief ascriptions, as well as potentially allowing a specificity of analysis that might not otherwise be present.

*[[This possibility seems to have been widely overlooked. For example, it is not mentioned in Schiffer's (1990) otherwise thorough survey of potential modes of presentation, even though primary intensions seem tailor-made to satisfy what Schiffer calls "Frege's constraint" on modes of presentation.]]

*[[Thus we are left with a framework for the semantics of belief ascription of the same general form (ascription of a propositional content plus constraint of a cognitive element) as the proposals described in Crimmins 1991, Richard 1990, and Schiffer 1992, although I have left the question of the logical form of these ascriptions open. Many of the insights of these and other philosophers on the semantics of belief ascription should be straightforwardly adaptable to the present framework.]]

(The fact that the constraints on primary intensions are only partial helps us to explain how Kripke's Pierre can believe that London is pretty and that London is not pretty without contradiction. What is going on is that Pierre's two concepts of London have different primary intensions (as we saw earlier), so that the beliefs' notional contents are compatible, but the primary intensions are both the *sort* that can make a "London" belief ascription true. Thus, what Kripke calls the "Principle of Non-Contradiction" is false: someone can rationally believe that *P* and believe that not-*P* as long as the beliefs involve two different primary intensions both of which satisfy the appropriate constraint.)

Given that notional content governs action, it follows that if belief ascriptions are to causally explain action, it must be in virtue of the notional content ascribed; the relational content ascribed is redundant to the explanation. To make this case more directly requires an examination of many specific examples wherein belief ascriptions are used to explain action, but the general point can be straightforwardly illustrated. One way to see the primacy of notional content is to consider beliefs involving empty names, such as "Santa Claus". These beliefs have no non-trivial relational content, and no non-trivial relational content is ascribed by the corresponding ascriptions, but ascription of beliefs about Santa Claus seem to function in *precisely* the same way in the explanation of action as do ascriptions of beliefs about real

people. We might explain young Karen's agitation on Christmas Eve in terms of her belief that Santa Claus is coming, that he will not fit down the chimney, and so on. Santa's non-existence and the corresponding absence of relational content make little difference to the success of such an explanation. What governs Karen's actions are her *notions* of Santa Claus; and what governs the success of the explanation is the notional content that these ascriptions ascribe to her. And this is a *typical* case of the role of belief ascriptions in explanation: even when non-trivial relational content is ascribed (as when the referent of the name exists), it makes little difference to the patterns of explanation.

In a very wide variety of cases in which content explains action, we can see that the explanation succeeds even if the relational content attributed is ignored. For instance, if we explain my opening the refrigerator in terms of my belief that there is water in the refrigerator and my desire for a glass of water, we never need to invoke the H₂O-involving relational content. The explanation gains sufficient purchase from the notional content ascribed alone - roughly, the content that there is some of the liquid with the appropriate properties in the refrigerator, and that I want some of that liquid, and so on.

It might be objected that there are cases in which the constraints on the notional content ascribed by a belief ascription are weak, so that relational content must be doing any explanatory work. I think that ascriptions putting weak constraints on notional content are rare, but assume they occur - perhaps an attribution of a belief about Smith constrains the relevant primary intension very little. Even so, if we look at how such an ascription can be used in explanatory practice, we see that notional content is still doing the real work. For example, perhaps we explain why Bev goes to the pub by saying that she wants to see Smith and believes that Smith is at the pub. Leaving aside constraints in the concepts of seeing, the pub, and so on, there is even a constraint on notional content implicit in the "Smith" attributions. For notice: it is implicit here that the two "Smith" concepts in Bev's thoughts have the same primary intensions. If her belief and her desire had very different notional content where "Smith" is concerned - perhaps she wants to see Batman and believes that Smith is at the pub, not knowing that Smith is Batman - the inference from those states to her action would fail. So there is a strong joint constraint on notional content: despite a lack of constraint on the individual beliefs, Bev is implicitly ascribed the belief that a person she wants to see is at the pub. It is *this* ascribed belief that is doing the real explanatory work, and this ascription clearly puts a heavy constraint on notional content. To make the case that all such examples can be similarly analyzed requires a detailed treatment, but this illustrates the general pattern.

It follows that the centrality of narrow content in the causation of action need not overthrow the role of belief ascriptions in explaining behavior, as some have suggested it should. At most we have shown that belief ascriptions are a somewhat rough-edged tool, due to the way they wrap both components of content into a single parcel, bringing the idle relational content into play alongside the notional content that does all the work.[*] But this should not surprise us; we cannot expect a folk theory to be maximally efficient.

*[[One might wonder why relational content is ascribed at all, given its explanatory idleness. I think there are two answers, both tied to language. First, given that we ascribe beliefs in the same language we use to describe the world, it is natural to do so in a world-involving way. When we use world-involving language to ascribe notional content, constraints on relational content come along for free. Second, relational content is important to understanding the success of communication and of collective action. When I tell you that I have a cold, you acquire a thought whose notional content is different from mine, but whose relational content is the same. Communication very frequently involves transmission of

relational content, and our collective cohesion (if not our individual actions) can often be understood in terms of shared relational content. But both of these points deserve a much more extensive development.]]

In moving from common-sense psychology toward a developed cognitive psychology, we might expect that the kind of content that is invoked will become more purely notional, and that relational content will be relegated to a secondary role or dropped entirely.[*] We might also expect that better tools will be developed to specify the notional contents of thoughts than the current rough-and-ready language of belief ascription. This might qualify as a revision of our folk notion of belief, emphasizing and refining the notional elements that are already present within it. But precisely because those elements are already present and playing a central role in our practices, such a development would fall well short of elimination.

*[[It is not implausible to argue that cognitive psychology is already most concerned with notional content rather than relational content, insofar as it is concerned with content at all. For example, the psychological literature on concepts seems to be largely concerned with how concepts are applied to the actual world, concentrating on something like the primary intensions of the concepts involved. See Smith and Medin 1981 and Patterson 1991.]]

6 Related Suggestions and Some Objections

There are various existing proposals to which the framework I have suggested is closely related. Perhaps the most obvious relationship is to Kaplan's (1989) two-dimensional framework for the semantics of demonstratives. Kaplan defines the "content" of a statement or term as a function from circumstances of evaluation into extension, and defines its "character" as a function from context of utterance into content. Kaplan's circumstances and contexts play the much same role as the uncentered and centered worlds used here. His content is just relational content as I have defined it, and something like a primary intension can be recovered from character by evaluating character in a centered world and evaluating the resulting content in the same world.

Because it is tied to language rather than thought, however, Kaplan's account has some marked differences with the account I have presented. On Kaplan's account, a name or a natural-kind term has the same content in any context, so that its character is relatively trivial. The English term "water" picks out H₂O in all contexts, for example. (Twin Earth "water" picks out XYZ, but it is a different word in a different language.) For similar reasons, co-referential names such as "Hesperus" and "Phosphorus" have identical character. As Kaplan notes, this means that his account does not have the resources to handle cognitive puzzles such as Frege's in general, although it works well in the case of indexicals and demonstratives. And because character often depends on reference, this does not yield an account of narrow content.

A generalization to thought has been made by Fodor (1987) (and in a related suggestion by White (1982)), who defines the content of a thought the usual way (as relational content), and defines the narrow content of a thought as a function from contexts to content. Unlike its linguistic counterpart, a "water" thought can have different contents in different contexts, so the results are more fine-grained than on Kaplan's account. The resulting formalism is similar to the present one, with the crucial difference that by making

relational content primary, narrow content becomes entirely derivative and its independent *semantic* function is lost. Fodor says as much:

Indeed, if you mean by content what can be semantically evaluated, then what my water-thoughts share with Twin "water"-thoughts isn't content. Narrow content is radically inexpressible, because it's only content *potentially*; it's what gets to be content when - and only when - it gets to be anchored. We can't - to put it in a nutshell - say what Twin thoughts have in common. This is because what can be said is ipso facto semantically evaluable; and what Twin-thoughts have in common is ipso facto not. (Fodor 1987, p. 50.)

Fodor and White nowhere suggest that narrow content has truth-conditions of its own. Despite the formal similarities, these accounts miss the crucial point that by making narrow content a function from context into *truth-value* (or from context into referent, for concepts), narrow content becomes semantically evaluable in its own right, along an independent semantic dimension. This small difference is entirely responsible for the explanatory power of the current framework. By allowing semantic relations such as implication between narrow contents, for example, it allows us to draw the direct link between narrow content and rationality. It is also this that allows us to analyze narrow content in terms of the worlds that the content endorses.

Another related suggestion that is canvassed occasionally (e.g. in Fodor (unpublished)) is the proposal that the narrow content of a concept such as "water" corresponds to a surface-level description of its referent, or a list of its apparent properties, such as "the clear, potable, thirst-quenching liquid found in the oceans and lakes". But as many have pointed out,[*] this strategy suffers from the problem that the resulting content is not narrow but wide. The content of a term like "liquid", for instance, may be bound to the empirical internal structure of liquids. We can construct a twin case, where liquids are replaced by superficially identical masses of fine grains of sand. Intuitively, my twin does not share even the content specified above with me, as his beliefs are not about *liquids* at all.

*[[E.g., LePore and Loewer 1986; Taylor 1989; White 1982.]]

It might be thought that this criticism applies to the present approach. After all, I have said that the primary intension of "water" picks out approximately the clear liquid with the properties mentioned. But importantly, this is only to get a rough handle on the intension for the purposes of illustration. The real narrow content is an abstract function from centered worlds into extensions, and is only imperfectly captured by a given description. We can only characterize the intension fully by detailed consideration of its value at specific worlds. As soon as we move to a summarizing description in language, imperfections are introduced, and the narrowness of the content is impurified. But the intension itself remains narrow; we should not mistake the linguistic description for the real thing.[*]

*[[This might suggest that notional content is "ineffable", but the real problem is that linguistic descriptions of intensions are usually read relationally, and it is hard to capture one concept's notional content in terms of others' relational content. To express notional contents, we might use a device such as square brackets to indicate that descriptions of intensions are to be read notionally (when the relevant linguistic expressions have an associated notional content, as many do). For example, we

might say that the notional content of my "water" concept is approximately the function that picks out [the dominant clear liquid in my environment] in a given centered world; that is, the referent of the former in an actual-world candidate is approximately the referent of the latter. This way all relational elements are removed. Of course this is characterizing a notional content in terms of another notional content, but it is hard to see why this is any more objectionable than characterizing a relational content in terms of another relational content. (Thanks to Frank Jackson for discussion on this point.))]

One would obtain a more closely related sort of "description" theory of notional content if one abstracted away from linguistic characterizations and regarded the relevant "descriptions" simply as properties that a referent must satisfy, or better, as relations to the thinker. If we speak merely of properties and relations, the linguistic contamination is avoided. Schiffer (1978) gives a description theory of this sort, on which there is irreducibly *de re* reference by a thinker to himself or herself, with reference to everything else is mediated by a property or relation. If we map the irreducible self-reference here to the appeal to centered worlds, and map the properties and relations to primary intensions, the resemblance between the accounts is clear, although Schiffer does not appeal to a two-dimensional framework, and addresses his proposal largely to the question of accounting for *de re* thought.

An even closer relationship to notional content as I have described it is present in Lewis's (1979) proposal that belief is a self-ascription of a property. The set of individuals satisfying a property corresponds directly to a class of centered worlds, as Lewis notes, and the relevant worlds are of the notional-world variety, so that for example "water" beliefs endorse individuals in H₂O-worlds and XYZ-worlds equally. Lewis (1994) argues that this sort of content is narrow and is primary in explanation. Lewis's account is one-dimensional, but he indicates various ways in which wide content can be seen to be derivative.

Something very like the present two-dimensional proposal has been canvassed by Stalnaker (1990), in a paper that is generally *critical* of narrow content. Stalnaker takes wide (relational) content to be primary, and defines the "diagonal" proposition associated with a thought as the function that evaluates a thought's wide content in its own context. He suggests that the diagonal proposition might be what advocates of narrow content have in mind. The diagonal proposition is of course very similar to the primary proposition, which I have suggested is the notional content of a thought.

In his brief but rich discussion, Stalnaker makes three objections to this construal of narrow content, all of which are worth addressing. The first is as follows.

First, this explanation makes clear that the determination of narrow content presupposes that we can identify a thought independently of its content. [...] It is not at all clear in the case of beliefs, intentions, and other states and attitudes that one can identify something that is the belief and intention in abstraction from its content, something about which we can ask, what would the content of *that* belief have been if *it* had been a belief I had on twin earth. (Stalnaker 1990, p. 135.)

This is surely misguided. If we could not fix reference to a thought by something other than its content, it would be impossible for us to have non-trivial knowledge of any thought's content; any knowledge that a

thought has a given content would be trivial, like our knowledge that Julius is the inventor of the zip. In practice, we identify mental states in all sorts of ways - as something like *this* belief, or the belief expressed by such-and-such an utterance - and then we notice how they behave under evaluation. We can evaluate a thought in worlds one way, and thus determine its relational content; and we can evaluate the thought in centered worlds another way, thus determining its notional content. The situation with respect to the two kinds of content are quite symmetrical. It is even possible for us to fix reference to a particular belief by wide content, and then investigate its notional content. For example, we can talk about "Pierre's belief that London is pretty", and ask how its truth-value and relational content might vary in different centered worlds.

Someone might argue that relational content is so central to a thought that we cannot even talk about how the thought's reference and relational content might vary if different centered worlds are considered as actual. For example, it might be held that any concept that could potentially refer to XYZ in a different centered world could not be *our* concept "water"; perhaps we are inadvertently importing our twins' *different* concept instead. But this is to trivialize the case against notional content, and also to lose contact with common sense. It is clear that there is at least *a* way of individuating concepts and thoughts so that we can speak of how their reference and truth-value might vary in this way depending on how the actual world turns out; the very fact that we have solid intuitions about how we would describe various potential actual-world scenarios indicates this. To ignore this fact is to ignore a central aspect of our conceptual structure. Even if there are other ways of individuating concepts that yield different results, the question of which of these is the one true "concept" is not a question that needs to be answered.

Stalnaker's second objection is the following:

Even if these questions receive satisfactory answers, it will remain true that the narrow content or realization conditions of a thought, explained as I have explained them, will be defined only relative to a very limited set of possible worlds: specifically, only for possible worlds containing the thought token. The realization conditions for the thought are satisfied in a given possible world if and only if the proposition expressed by the thought *in that possible world* is true. Suppose the thought I am thinking could be naturally expressed as follows: "There is a hole in the ozone layer". If I focus on the wide content, then I can consider not only whether this thought is true in the actual world, where I am thinking it, but also whether it would be true if, say, human beings had never existed to pollute the atmosphere. But we can't evaluate the realization conditions in such a possible world, since the thought won't exist there. We might try to extend the procedure to apply to possible worlds not containing the thought by asking a counterfactual question about the thought: what *would* the content of the thought be if it were thought there? This would work in some cases, but it is likely to leave a large amount of indeterminacy in the narrow content of our thoughts. Consider a version of another one of our stories: Bert, who is a little confused about arthritis, thinks his father has this disease in his thigh. [...] Now consider a possible world in which both the language and the state of medical knowledge are quite different. In this world, there is no word like 'arthritis', in Bert's language, and Bert has no thoughts about his father's health. Are the realization conditions for Bert's actual thought satisfied in

this counterfactual world? Bert didn't have the belief there, but if he had, would it have been true? So far as I can see, there is no way to tell. (Stalnaker 1990, pp. 136-7.)

In language of this paper, this comes to the objection that the primary intension of a concept and the notional content of a thought will be indeterminate when evaluated at many centered worlds. There are a number of things to say to this. First, as Stalnaker observes, the truth-value of a thought in a centered world will be most straightforward when the centered world contains the thought at its center, but we have seen that this is not required. For example, a world with an artichoke at its center is precisely the sort of actual-world candidate that is endorsed by my thought "I am an artichoke", even if the artichoke is not thinking. In these cases we can retain the thought from the *real* actual world and simultaneously ask evaluate its truth-value in other actual-world candidates without any loss of coherence.[*]

We can handle Stalnaker's second example in a similar way. If Bert's actual world turned out to be a world in which no-one used any terms like "arthritis", then to what would his semantically deferential concept "arthritis" refer? It seems most natural to say that it would refer to nothing at all. This would be a world about which his *background assumptions* would be very mistaken. The use of many concepts depends on background assumptions, as Russell's example "The present King of France" shows. If those assumptions do not hold in the actual world, it is reasonable to say that thoughts involving these concepts lack truth-value. When Bert uses a concept such as "arthritis" with semantic deference, he makes a background assumption that people in the community around him use the word or something like it. If they do not, then all bets are off.

In any case, Bert's thoughts involving "arthritis" still have a significant notional content. They rule in and rule out certain centered worlds in which the background assumptions are satisfied. They take no stand on worlds in which the background assumptions are not satisfied, but that was never their purpose. For the usual explanatory purposes, the inclusion and exclusion of worlds relative to those assumptions will suffice. Centered worlds in which those assumptions fail will usually be ruled out by other beliefs, such as Bert's metalinguistic beliefs in these very assumptions (for example, his belief that the term "arthritis" is used by the community around him), so that his attitude toward these worlds can be handled in other ways. For the usual explanatory purposes, however, a belief ascriber will be concerned with a thinker's attitude toward worlds that are closer to home, so it is not a problem that notional content and primary intensions are only partial functions.

Another class of concepts whose primary intensions have a limited domain of definition are perceptual demonstratives. When I think something like "That is pretty", the referent of my demonstrative is often picked out (very roughly) as the cause of such-and-such experience.[*] In a centered world in which a pretty thing is appropriately responsible for the relevant experience at the center, the thought will be true; in a world in which an ugly thing is responsible, the thought is false; but in a world in which there is no appropriate experience at the center, the thought may lack truth-value. In this case fixation of reference relies on the fact that such an experience is present, just as fixation on reference in the "arthritis" case relies on the existence of a surrounding community using the term. Again, it is not a problem that worlds are ruled in and ruled out only relative to a background condition. If for some purpose we wish to invoke the thinker's attitudes to other worlds (a world where the experience is absent, for instance), we can rely

on the notional content of other thoughts in the vicinity, such as "The object in front of me is pretty".

*[[This raises another subtlety: to capture the content of certain perceptual demonstratives, one may need to build in a "marked" experience to the center of the class of actual-world candidates, as one builds in a marked individual and time. This is reflected in characterizations such as "cause of *this* experience", which may be irreducible in a way analogous to "next to *me*". (This parallels Russell's idea that reference to our experiences and to ourselves is irreducibly direct.) Building in this property of a center may secure reference to otherwise indistinguishable experiences and their perceptual objects, as with (perhaps) a speckle in a large field, or one of the red spots in Austin's (1990) "Two Tubes" puzzle (to which the present framework then provides a solution). In certain cases, centers may also require a marked thought ("this very thought"), further marked experiences, and perhaps marked orientations (e.g. to distinguish left and right). In general, what is built into a center corresponds to the things to which one can intuitively make "unmediated" reference: one's world, oneself, the present moment, the current thought, and perhaps certain experiences and orientations (though some of these might be seen as derivative on others). This is a matter I hope to explore elsewhere.]]

The precise domain of definition for the notional contents of particular thoughts is an important issue, for which there will be different answers depending on the thoughts in question. Some concepts, such as pure indexicals ("I", "here", and "now"), have primary intensions that are defined across all centered worlds. Pure non-indexicals - concepts whose referent at a centered world does not depend on the position of the center - will also have a wide domain of definition. It is concepts in between, for which the existence of a referent depends on background assumptions about how things are around the center of a world - as with semantically deferential concepts and certain demonstratives - that will have a restricted domain. This is a deep issue that deserves a detailed treatment of its own, but for now it suffices to note it is no objection to this account to note that the relevant functions are often partial.

It is occasionally suggested that there is something odd about the idea of evaluating a thought in an actual-world candidate that does not contain a copy of the thought itself. After all, we are considering the centered world as actual, but still retaining the thought from the actual world (the *real* actual world). But this seems simply to reflect how our thoughts work: they can endorse certain centered worlds as potential environments (that is, as actual-world candidates) without requiring those worlds to contain specific thoughts at the center. The thought "I am comatose" provides a clear example, as does a thought such as "Julius was a genius". Certainly it seems mistaken to require that a thinker's notional world-model contains a perfect copy of all of his or her thoughts.

This reflects itself in the fact that we have solid intuitions about how to (notionally) describe actual-world candidates even when those candidates do not contain the relevant thoughts. Consider a centered world in which I am at the center thinking about watermelons, and in which the inventor of the zip is not the actual-world inventor (who died at birth here) but Albert Einstein (who is as smart as ever). Then I can say: if that world turned out to be actual, it would turn out that Julius was a genius (using the "turns out" locution in the usual notional way). Correspondingly, this world is among the actual-world candidates that my thought "Julius was a genius" endorses. (Of course it is not among the worlds that the *relational* content of my thought endorses.) The very fact that we have these solid intuitions is evidence that something is going on. Any semantic theory is grounded in the evidence of our intuitions, so these basic data should not be ignored.

The alternative would be to require a copy of the thought or the concept at the center of every world, but this would be to greatly limit the value of the framework in accounting for the contents of our thoughts, as it ignores the fact that our thoughts are involved in taking attitudes to notional worlds in which the thoughts are absent. This would also raise the problem of just what we would have to "hold constant" in copying a thought between centered worlds: a problem raised for Fodor's account by Block (1991) and Stalnaker (1991), but which the current framework avoids entirely. It seems much more satisfactory not to involve a thought itself in its conditions of application, even to actual-world candidates.[*]

*[[Of course, the application-conditions of *some* thoughts will involve the presence of a thought (although usually not presence of *the* thought). In these cases we can build in the presence of a thought as a background condition. "This thought" provides an example: it is a background condition that there is a "marked" thought at the center. Perhaps a semantically deferential concept such as my concept "vanadium" gains its link to the community by picking out whatever is causally responsible for my own use of the concept; if so, it may be a background condition that there is a marked concept labeled "vanadium" at the center. In these cases, evaluation in centered worlds in which relevant concepts are not present at the center will be indeterminate. As before, if we want to explain our attitude to those worlds, we will have to use other concepts.]]

Finally, Stalnaker suggests:

The source of all these problems with this procedure for defining narrow content is the fact that according to this account, narrow content is derivative from (actual and possible) wide content. The explanation of realization conditions takes for granted that somehow, the external and internal facts about a mental event or state that we have identified determine a content (in the ordinary sense) for that state. Then from the actual wide content, together with the facts about what the wide content would be under various counterfactual circumstances, we extract the narrow content. Narrow content, as defined by this procedure, presupposes rather than explains wide content. (Stalnaker 1990, p. 137.)

This, it seems to me, is false and importantly so. Perhaps narrow content is derivative on wide content when it is indirectly defined by Stalnaker's "diagonal" construction, but as I have characterized notional content, it is in no way derivative. There are two basic notions - what it is for a thought to be true *in* a world and what it is for a thought to be true *of* a world - and neither of these is obviously derivative on the other. Only if we have already adopted a framework in which relational content is primary will we determine the notional content of a thought by first determining its relational content in centered world and then evaluating that relational content in the same world. It is in many ways more natural to determine notional content directly, and to directly evaluate the truth of a thought in a world.

Indeed, in many ways relational content seems to be derivative on notional content. We can see this most clearly by looking at concepts rather than complete thoughts. We have already seen that the secondary intension of most concepts generally depends on the primary intension. In some cases it is a near-copy of the primary intension, as for simple descriptive concepts, and in other cases it is determined by rigidifying the actual-world evaluation of the primary intension. By contrast, we can tell the entire story about the primary intension without ever involving the secondary intension. It therefore seems that if either intension is to be fundamental, it is the primary intension. If we see notional and relational content as

composed from primary and secondary intensions respectively, and if we see secondary intensions as derivative on primary intensions, then it seems that narrow content may explain wide content rather than *vice versa*.[*]

*[[This suggests that in order to have a naturalistic theory of wide content, for example, we will probably need a naturalistic theory of notional content first. Arguably, the reason that contemporary causal theories of content have been unsuccessful is that all these attempt to account for relational content without first accounting for notional content.]]

It is unclear why the assumption that relational content is primary is so common in the recent literature. One reason is no doubt a concern with language: relational content is often more stably associated with a term than is notional content. Perhaps a second reason is an aversion to indexical accounts of content, grounded in Frege's insistence that the truth-value of a thought cannot depend on a thinker's location in a world. A third reason is that the structure of relational content is more naturally described in language than that of notional content, making notional content more difficult to discuss.[*]

*[[A vulgar psychohistory: since Frege it has been recognized that intensions are central to the analysis of content, to deal with cognitive puzzles, the fixation of reference, and the like. For a long time, the intensionality of reference-fixing was not clearly distinguished from that of modal contexts, so it was natural to characterize the relevant intensions with modal contexts such as "what is water in *w*?". After Kripke separated the two sorts of intensionality, philosophers continued to concentrate on the latter sort, even though it is ill-suited to play the main roles that intensionality was introduced into semantic theories to play. The deepest sort of intensionality lies in the fixation of reference; the sort that governs description of counterfactual scenarios is much less fundamental.]]

Whatever the reasons, the emphasis on relational content has been misplaced. In many ways, notional content is more central to thought than relational content: it is cognitively accessible, it governs the cognitive and rational relations between thoughts, and it governs behavior. But we need not dismiss relational content entirely. Its utility is still an open question, and it may have a central role at least in a theory of communication. The most satisfactory overall picture will emerge if we allow the two kinds of content to co-exist in a two-dimensional conception of meaning.

6 Conclusion

What of the six puzzles at the start? To summarize:

(1) A thought's content decomposes into notional and relational content, given by its associated primary and secondary propositions respectively. My twin and I differ in the relational contents of our thoughts, but our notional contents are the same.

(2) The thoughts that Hesperus is Hesperus and that Hesperus is Phosphorus have the same associated secondary proposition but distinct associated primary propositions, as the "Hesperus" and "Phosphorus" concepts have different primary intensions. The triviality of the former does not imply the triviality of the latter, as it is notional content that governs rational relations.

(3) Pierre's two beliefs have contradictory secondary propositions but compatible primary propositions, as his two concepts of London have different primary intensions, although they are both the sort of intension that can make "London"-involving belief ascriptions true. It is only contradictory primary propositions that imply irrationality.

(4) The essential indexicality of belief reflects the fact that notional content, not relational content, governs action, and that notional content, unlike relational content, is a centered proposition.

(5) The modes of presentation central to a theory of belief ascription are primary intensions and primary propositions. Belief ascriptions constrain not just the relational content but the notional content of the believer.

(6) Instances of the contingent *a priori* have a necessary primary proposition but a contingent secondary proposition. It is notional content that constrains one's world-model, so a contingent secondary proposition does not indicate a cognitive achievement.

There are many problems about the contents of thought that are not resolved by this framework. Most notable are problems with hyperintensionality: how do we accommodate cognitive differences between thoughts that have the same notional and relational contents, such as beliefs in different mathematical truths? That may require a new, refined kind of content, perhaps one that lies behind notional content and determines it in turn. There is also much work to be done in providing a full account of the semantics of belief ascription; and I have left the project of providing a naturalistic reduction of this sort of content almost entirely untouched. Some of these matters are likely to be much more difficult than the puzzles at issue in this paper, but the two-dimensional framework at least clarifies the lay of the land.

Acknowledgments

Thanks to Ned Block, Frank Jackson, David Kaplan, David Lewis, Andrew Milne, Gideon Rosen, Jerry Seligman, and Robert Stalnaker, to audiences at Cornell University, Princeton University, Rice University, UCLA, University of Memphis, and Washington University, and to two referees for *Mind*.

References

Austin, D.F. 1990. *What's the Meaning of "This"?* Ithaca, NY: Cornell University Press.

Block, N. 1988. Functional Role and Truth-Conditions. *Proceedings of the Aristotelian Society* 61:157-81.

Block, N. 1991. What Narrow Content is Not. In (B. Loewer & G. Rey, eds.) *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

- Burge, T. 1979. Individualism and the Mental. *Midwest Studies in Psychology* 4: 73-122.
- Chalmers, D.J. 1996. *The Conscious Mind*. MIT Press.
- Clark, A. & Chalmers, D.J. 1995. The Extended Mind. Philosophy/Neuroscience/Psychology Technical Report 95-21, Washington University.
- Crimmins, M. 1991. *Talk about Beliefs*. Cambridge, MA: MIT Press.
- Davies, M.K. & Humberstone, I.L. 1980. Two Notions of Necessity. *Philosophical Studies* 38:1-30.
- Dennett, D.C. 1981. Beyond Belief. In (A. Woodfield, ed.) *Thought and Object*. Oxford: Oxford University Press.
- Evans, G. 1979. Reference and Contingency. *The Monist* 62:161-89.
- Fodor, J. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.
- Fodor, J. 1991. A Modal Argument for Narrow Content. *Journal of Philosophy* 88:5-26.
- Frege, G. 1892. Über Sinn und Bedeutung. Translated in (P. Geach & M. Black, eds.) *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Blackwell, 1952.
- Kaplan, D. 1989. Demonstratives. In (J. Almog, J. Perry, and H. Wettstein, eds.) *Themes from Kaplan*. Oxford: Oxford University Press.
- Kripke, S.A. 1979. A Puzzle about Belief. In (A. Margalit, ed.) *Meaning and Use*. Dordrecht: D. Reidel.
- Kripke, S.A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
- LePore, E. & Loewer, B. 1986. Solipsistic Semantics. *Midwest Studies in Philosophy* 10:595-614
- Lewis, D. 1979. Attitudes *De Dicto* and *De Se*. *Philosophical Review* 88: 513-43.
- Lewis, D. 1994. Reduction of Mind. In (S. Guttenplan, ed.) *Companion to the Philosophy of Mind*. Oxford: Blackwell.
- Loar, B. 1988. Social Content and Psychological Content. In (R.H. Grimm and D.D. Merrill, eds.) *Contents of Thought*. Tucson: University of Arizona Press.
- Patterson, S. 1991. Individualism and Semantic Development. *Philosophy of Science* 58:15-35.

- Perry, J. 1977. Frege on Demonstratives. *Philosophical Review* 86: 474-97.
- Perry, J. 1979. The Problem of the Essential Indexical. *Nous* 13:3-21
- Putnam, H. 1975. The Meaning of 'Meaning'. In (K. Gunderson, ed.) *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press.
- Putnam, H. 1981. *Reason, Truth, and History*. Cambridge: Cambridge University Press.
- Quine, W.V. (1968). Propositional Objects. *Critica* 2(5):3-22.
- Richard, M. 1990. *Propositional Attitudes: An Essay on Thoughts and How we Ascribe Them*. Cambridge: Cambridge University Press.
- Schiffer, S. 1978. The Basis of Reference. *Erkenntnis* 13: 171-206.
- Schiffer, S. 1990. The Mode-of-Presentation Problem. In (C.A. Anderson & J. Owens, eds.) *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*. Stanford: CSLI Press.
- Schiffer, S. 1992. Belief Ascription. *Journal of Philosophy* 87:602-14.
- Smith, E.E. & Medin, D.L. 1981. *Categories and Concepts*. Cambridge, MA: Harvard University Press.
- Stalnaker, R. 1978. Assertion. In (P. Cole, ed.) *Syntax and Semantics: Pragmatics, Vol. 9*. New York: Academic Press
- Stalnaker, R. 1990. Narrow Content. In (C.A. Anderson and J. Owens, eds.) *Propositional Attitudes*. Stanford: Center for the Study of Language and Information.
- Stalnaker, R. 1991. How to Do Semantics For the Language of Thought. In (B. Loewer & G. Rey, eds.) *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.
- Taylor, K. 1989. Narrow Content Functionalism and the Mind-Body Problem. *Nous* 23:355-72
- White, S. 1982. Partial Character and the Language of Thought. *Pacific Philosophical Quarterly* 63:347-65.
- White, S. 1991. Narrow Content and Narrow Interpretation. In *The Unity of the Self*. Cambridge, MA: MIT Press.

Philosophy of Mind: Classical and Contemporary Readings

Edited by [David J. Chalmers](#)

A new anthology of articles in the philosophy of mind, *Philosophy of Mind: Classical and Contemporary Readings* will be published by Oxford University Press in July 2002. This is intended as a comprehensive collection that can be used in university courses at all levels (introductory, advanced undergraduate, graduate). The volume concentrates on foundational issues about the metaphysics of mind, consciousness, and mental content. (There will eventually be a companion volume on the philosophy of cognitive science.) I am told that the book (paperback, around 700 pages, \$44.95) will be published on July 19, in plenty of time to be used for courses in fall 2002.

The table of contents for the volume is below. By publication date, I will flesh out this web page with additional information and links useful for readers of the book.

TABLE OF CONTENTS

Preface

PART I: FOUNDATIONS

Introduction

A. Dualism

1. Rene Descartes, *Meditations on First Philosophy* (II and VI).
2. Rene Descartes, *Passions of the soul* (excerpt).
3. Thomas H. Huxley, *On the hypothesis that animals are automata, and its history* (excerpt).
4. Raymond M. Smullyan, *An unfortunate dualist*.

B. Behaviorism

5. Gilbert Ryle, *Descartes' myth*.

6. Rudolf Carnap, Psychology in physical language (excerpt).

7. Hilary Putnam, Brains and behavior.

C. The identity theory

8. U.T. Place, Is consciousness a brain process?

9. J. J. C. Smart, Sensations and brain processes.

10. Herbert Feigl, The "mental" and the "physical" (excerpt).

D. Functionalism

11. Hilary Putnam, The nature of mental states.

12. David M. Armstrong, The causal theory of the mind.

13. David Lewis, Psychophysical and theoretical identifications.

14. Ned Block, Troubles with functionalism (excerpt).

15. Martine Nida-Rümelin, Pseudonormal vision: An actual case of qualia inversion?

E. Other psychophysical relations

16. C. D. Broad, Mechanism and its alternatives (excerpt).

17. Donald Davidson, Mental events.

18. Jerry A. Fodor, Special sciences.

19. Jaegwon Kim, Multiple realization and the metaphysics of reduction.

20. Terence Horgan, From supervenience to superdupervenience: Meeting the demands of a material world (excerpt).

21. Frank Jackson, Finding the mind in the natural world.

F. Mental causation

22. Jaegwon Kim, The many problems of mental causation (excerpt).

23. Stephen Yablo, Mental causation.

PART II. CONSCIOUSNESS

Introduction

A. General

24. Ned Block, Concepts of consciousness.

25. Thomas Nagel, What is it like to be a bat

26. Daniel C. Dennett, Quining qualia.

27. David J. Chalmers, Consciousness and its place in nature.

B. The knowledge argument

28. Frank Jackson, Epiphenomenal qualia.

29. David Lewis, What experience teaches.

30. Brian Loar, Phenomenal states (second version).

31. Daniel Stoljar, Two conceptions of the physical.

C. Modal arguments

32. Saul A. Kripke, Naming and necessity (excerpt).

33. Christopher S. Hill, Imaginability, conceivability, possibility, and the mind-body problem (excerpt).

34. Grover Maxwell, Rigid designators and mind-brain identity (excerpt).

D. The explanatory gap

35. Joseph Levine, Materialism and qualia: The explanatory gap.

36. Paul M. Churchland, The rediscovery of light.

37. Ned Block and Robert Stalnaker, Conceptual analysis, dualism, and the explanatory gap.

38. Colin McGinn, Can we solve the mind-body problem?

E. Higher-order thought and representationalism

39. David Rosenthal, Explaining consciousness.

40. Fred Dretske, Conscious experience.

41. Christopher Peacocke, Sensation and the content of experience: A distinction.

42. Michael Tye, Visual qualia and visual content revisited.

43. Sydney Shoemaker, Introspection and phenomenal character.

PART III. CONTENT

Introduction

A. The nature of intentionality

44. Franz Brentano, Mental and physical phenomena (excerpt).

45. Roderick M. Chisholm, "Intentional inexistence".

46. Fred Dretske, A recipe for thought.

47. Ruth Millikan, Biosemantics.

48. Robert M. Brandom, Reasoning and representing.

49. Terence Horgan & John M. Tienson, The intentionality of phenomenology and the phenomenology of intentionality.

B. Propositional attitudes

50. Wilfrid Sellars, Empiricism and the philosophy of mind (excerpt).

- 51. Jerry A. Fodor, Propositional attitudes.
- 52. Daniel C. Dennett, True believers: The intentional strategy and why it works.
- 53. Paul M. Churchland, Eliminative materialism and the propositional attitudes.

C. Internalism and externalism

- 54. Hilary Putnam, The meaning of "meaning" (excerpt).
- 55. Tyler Burge, Individualism and the mental (excerpt).
- 56. David J. Chalmers, The components of content (revised version).
- 57. Michael McKinsey, Anti-individualism and privileged access.
- 58. Anthony Brueckner, What an anti-individualist knows a priori.
- 59. Andy Clark & David J. Chalmers, The extended mind.

PART IV: MISCELLANEOUS

Introduction

- 60. Derek Parfit, Reductionism and personal identity.
- 61. A. J. Ayer, Freedom and necessity.
- 62. Bertrand Russell, Analogy.
- 63. John R. Searle, Can computers think?

Does Conceivability Entail Possibility?

[David J. Chalmers](#)

**Department of Philosophy
University of Arizona
Tucson, AZ 85721.**

chalmers@arizona.edu

[[To appear in *Imagination, Conceivability, and Possibility*, edited by Tamar Gendler and John Hawthorne, forthcoming from Oxford University Press.]]

There is a long tradition in philosophy of using a priori methods to draw conclusions about what is possible and what is necessary, and often in turn to draw conclusions about matters of substantive metaphysics. Arguments like this typically have three steps: first an epistemic claim (about what can be known or conceived), from there to a modal claim (about what is possible or necessary), and from there to a metaphysical claim (about the nature of things in the world).

We find this structure in many different areas of philosophy: in arguments about whether the mental is reducible to the physical (or vice versa), about whether causation and laws are reducible to regularities in nature, about whether knowledge is identical to justified true belief, and so on. Many arguments in these domains first seek to establish an epistemic gap between two phenomena (e.g. that we can know or conceive of one without the other), argue from there to a modal gap (e.g. that it is possible that one could exist without the other), and step from there to a metaphysical gap (e.g. that one is not reducible to the other).

Here, I will mostly be concerned with the second step: the bridge between the epistemic and modal domains. The most popular bridge here is the method of conceivability. One argues that some state of affairs is conceivable, and from there it is concluded that this state of affairs is possible. Here, the kind of possibility at issue is *metaphysical* possibility, as opposed to physical possibility, natural possibility, and other sorts of possibility. Metaphysical conclusions turn most directly on matters of metaphysical possibility: if one domain is reducible to another, the facts about the second should metaphysically necessitate the facts about the first. So it is metaphysical possibility that is relevant in the three-step argument above. And there is at least some plausibility in the idea that conceivability can act as a guide to metaphysical possibility.

By contrast, it is very implausible that conceivability entails physical or natural possibility. For example, it seems conceivable that an object could travel faster than a billion meters per second, but this may be physically and naturally impossible, because it contradicts the laws of physics and the laws of nature. This case may be metaphysically possible, however, since there might well be metaphysically possible worlds

with different laws. If we invoke an intuitive conception of a metaphysically possible world as a world that God might have created, if he had so chosen: it seems that God could have created a world in which an object traveled faster than a billion meters per second. So in this case, although conceivability does not mirror natural possibility, it may well mirror metaphysical possibility.

In recent years, conceivability arguments have faced considerable opposition. Many philosophers hold that the step from conceivability to metaphysical possibility has been shown to be invalid, not least due to a number of apparent counterexamples. Things are complicated by the fact that the term "conceivability" means many different things to different people. In some senses of the term, an entailment from conceivability to possibility is out of the question; in other senses things are not so clear. To properly assess this matter, we first have to clarify what is meant by "conceivability".

Here I will isolate three dimensions of difference between notions of conceivability: *prima facie* vs. ideal conceivability, primary vs. secondary conceivability, and positive vs. negative conceivability. These distinctions are largely independent of each other, so there may be up to eight sorts of conceivability in the vicinity: *prima facie* primary positive conceivability, and so on. By making these distinctions, I think at least one plausible and defensible conceivability-possibility thesis can be formulated, free of any clear counterexamples.

As I will be using the term here, conceivability is a property of statements, and the conceivability of a statement is in many cases relative to a speaker or thinker. I think that conceivability is more deeply a property of propositions, but I will not talk that way here, since many philosophers have theoretical views about propositions that can confuse these issues. For a statement *S*, we will have eight or so ways of disambiguating the claim that *S* is conceivable for a given subject. I will first give rough characterizations of the various dimensions of difference. Then I will examine various specific notions of conceivability that result, and address the question of the extent to which these notions of conceivability support an entailment from the conceivability of *S* to the possibility of *S*. For ease of discussion, I will use sentence symbols such as '*S*' loosely, allowing context to disambiguate whether the corresponding sentence is being used or mentioned.

PRIMA FACIE vs. IDEAL CONCEIVABILITY: *S* is *prima facie* conceivable for a subject when *S* is conceivable for that subject on first appearances. That is, after some consideration the subject finds that *S* passes the tests that are criterial for conceivability. The specific criteria will depend on a substantive notion of conceivability, as outlined in the discussion of the remaining dimensions of conceivability, to remove the apparent circularity. For example, one substantive notion of conceivability (a version of negative conceivability) holds that *S* is conceivable if no contradiction is detectable in the hypothesis expressed by *S*. Under this notion, *S* will be *prima facie* conceivable for a subject when that subject cannot (after consideration) detect any contradiction in the hypothesis expressed by *S*.

S is ideally conceivable when *S* is conceivable on ideal rational reflection. It sometimes happens that *S* is *prima facie* conceivable to a subject, but that this *prima facie* conceivability is undermined by further reflection showing that the tests that are criterial for conceivability are not in fact passed. In this case, *S* is not ideally conceivable. Given the substantive notion of (negative) conceivability above, for example, *S*

will be ideally conceivable when ideal rational reflection detects no contradiction in the hypothesis expressed by S ; or equivalently, when $\sim S$ is not a priori.

An example is provided by any mathematical statement M whose truth-value is currently unknown, but which will later be proved to be true. Here $\sim M$ is prima facie conceivable in the sense above (i.e. prima facie negatively conceivable) at least for current subjects, but it is not ideally conceivable, as ideal reflection will rule out $\sim M$ a priori.

The notion of ideal rational reflection remains to be clarified. One could try to define ideal conceivability in terms of the capacities of an ideal reasoner - a reasoner free of all contingent cognitive limitations. Using this notion, we could say that S is ideally conceivable if an ideal reasoner would find it to pass the relevant tests (if an ideal reasoner could not rule out the hypothesis expressed by S a priori, for example). A strategy like this is taken by Menzies (1998). One trouble is that it is not obvious that an ideal reasoner is possible or coherent. For example, it may be that for every possible reasoner, there is a more sophisticated possible reasoner. Alternatively, one can dispense with the notion of an ideal reasoner, and simply invoke the notion of undefeatability by better reasoning. Given this notion, we can say that S is ideally conceivable when there is a possible subject for whom S is prima facie conceivable, with justification that is undefeatable by better reasoning. The idea is that when prima facie conceivability falls short of ideal conceivability, then the claim that the relevant tests are passed will either be unjustified, or the justification will be defeatable by further reasoning. For ideal conceivability, one needs justification that cannot be rationally defeated.

I will not try to give a substantive characterization of what good reasoning consists in, or of what counts as a cognitive limitation to be idealized away from. I suspect that any such attempt would end up being open-ended and incomplete. In general, my approach in this paper is to take certain rational notions as primitive, and to what sort of connection to modal notions emerges. In this case, I am simply appealing to our intuitive grasp of notions of reasoning, and of when one reasoning process defeats another. I note that the notion of undefeatability here invoked here is also implicit in our concept of knowledge: it is generally held that if one's justification for a belief that P is defeatable by better reasoning, then one does not know that P . So the notion of conceivability is not obviously worse off than the concept of knowledge.

There is also a fairly direct parallel between the idealization present in the notion of ideal conceivability and that present in the familiar notion of apriority. If I cannot know that P independent of experience, but another less limited being could do so, then it is a priori that P . And if I believe that P , but the justification for my belief is defeatable by better reasoning, then it is not a priori that P (unless there is another undefeatable justification). So the notion of apriority idealizes away from cognitive limitations in much the same way as the notion of ideal conceivability. This is not to say that either of these idealizations are perfectly clear, but at least the idealization is a familiar one. And in practice, the idealizations are easy to apply. We will see that there are certain difficult cases on the far end of idealization where things get tricky; but dealing with such cases may allow us to further clarify the idealization.

There are a couple of things that should be clarified in advance, however. First, it is important that "better reasoning" about conceivability not be defined even in part as reasoning that better tracks possibility.

Such a criterion would trivialize the link between ideal conceivability and possibility. Fortunately there is no reason to expect that such a criterion will come into play, at least on most of the substantive notions of conceivability we will be considering. Where conceivability is defined in terms of what is ruled out a priori, for example, we have an entirely independent grounding for the notion. Only if conceivability is directly defined in terms of possibility - perhaps as what a subject judges to be possible - will there be a danger of triviality.

Second, in most cases (with an exception to be discussed later), the reasoning in question is restricted to a priori reasoning, and the further reasoning involved in the idealization will remain within the a priori domain. Sometimes this will be an automatic consequence of a given notion of conceivability (e.g. the negative notion of conceivability above), and sometimes it can be seen as a stipulation. Either way, this restriction is important if this issue is to shed light on the issue of a priori access to modality.

POSITIVE vs. NEGATIVE CONCEIVABILITY: Negative notions of conceivability hold that S is conceivable when S is not *ruled out*. For example, a popular sense of "conceivable" in common usage holds roughly that S is conceivable when it is not ruled out by what one knows, or by what one believes. I will set this popular usage aside as tangential to our main purposes here: philosophers are usually concerned with senses in which S can be conceivable even when one knows that S is not actually the case. More relevant notions of negative conceivability can be obtained by constraining the ways in which S might be ruled out.

The central sort of negative conceivability holds that S is negatively conceivable when S is not ruled out a priori, or when there is no (apparent) contradiction in S. One can disambiguate the notion depending by applying the distinction between *prima facie* and ideal conceivability, as above. We can say that S is *prima facie* negatively conceivable for a subject when that subject, after consideration, cannot rule out S on a priori grounds. And we can say that S is ideally negatively conceivable when it is not a priori that $\sim S$.

One subtlety concerns cases of indeterminacy. For some S (perhaps statements that are not truth-evaluable, or some statements involving vague predicates), it may be a priori that it is indeterminate whether S. If so, it is not a priori that $\sim S$. In such a case, is S negatively conceivable? For various reasons, it seems best to say that it is not: in these cases, the possibility that S is not truly left open. To handle such cases, one can say that S is negatively conceivable when $\det(S)$ cannot be ruled out, and that S is ideally negatively conceivable when it is not a priori that $\sim \det(S)$. Here " $\det S$ " expresses the claim that S is determinately the case, and " $\sim \det(S)$ " expresses the claim that S is false or indeterminate. (In other frameworks for dealing with indeterminacy, one can adopt a corresponding definition.) In the case of a priori indeterminacy above, it will be a priori that $\sim \det(S)$, so S will not be ideally negatively conceivable.

Positive notions of conceivability require that one can form some sort of positive conception of a situation in which S is the case. One can place the varieties of positive conceivability under the broad rubric of *imagination*: to positively conceive of a situation is to in some sense imagine a specific configuration of objects and properties. It is common to imagine situations in considerable detail, and this imagination is often accompanied by interpretation and reasoning. When one imagines a situation and reasons about it, the object of one's imagination is often revealed as a situation in which S is this case, for some S. When

this is the case, we can say that the imagined situation *verifies* S, and that one has *imagined that* S. Overall, we can say that S is positively conceivable when one can imagine that S: that is, when one can imagine a situation that verifies S. (This definition, and the following discussion, is indebted to the discussion of conceivability in Yablo 1993.)

Different notions of conceivability correspond to different notions of imagination. One such notion is tied to *perceptual imagination*. A subject perceptually imagines that S when the subject has a perceptual mental image that represents S as being the case. This happens roughly when the image relevantly resembles a perceptual experience that represents S as being the case (see Gendler/Hawthorne, this volume[?]). For example, one can perceptually imagine that a pig flies by forming a visual image of a flying pig, where this can be understood as an image that relevantly resembles a visual experience as of a flying pig.

Perceptually imagining that P differs from supposing that P, or from entertaining the proposition that P, in that it involves not just an attitude toward P, but toward some specific situation that stands in a certain relationship to P. To perceptually imagine that pigs fly, we form a mental image that represents a specific situation (one with a certain configuration of animals), and we take this to be a situation in which pigs fly. Here, we can say that the imagined situation verifies "pigs fly". More generally, one can say that when one perceptually imagine that P, one perceptually imagines a situation that verifies P. Unlike entertaining or supposing that P, the phenomenology of perceptually imagining that P has a mediated objectual character, with an attitude toward an intermediate mental object (here, an imagined situation) playing a crucial role. This objectual character (noted by Yablo 1993) is distinctive of positive conceivability.

This objectual character is present in cases of imagination that are not grounded in imagery. There is a sense in which we can imagine situations that do not seem to be potential contents of perceptual experiences. One can imagine situations beyond the scale of perception: e.g. molecules of H₂O, or Germany winning the Second World War. One can imagine situations that are unperceivable in principle: e.g. the existence of an invisible being that leaves no trace on perception. And one can imagine pairs of situations that are perceptually indistinguishable: e.g. the situations postulated by two scientific hypotheses that make the same empirical predictions, or arguably the existence of a conscious being and its zombie twin (an unconscious physically identical duplicate).

In these cases, we do not form a perceptual image that represents S. Nevertheless, we do more than merely suppose that S, or entertain the hypothesis that S. Our relation to S has a mediated objectual character that is analogous to that found in the case of perceptual imaginability. In this case, we have an intuition of (or as of) a *world* in which S, or at least of (or as of) a situation in which S, where a situation is (roughly) a configuration of objects and properties within a world. We might say that in these cases, one can *modally imagine* that P. One modally imagines that P if one modally imagines a world that verifies P, or a situation that verifies P. Modal imagination goes beyond perceptual imagination, for the reasons above, but it shares with perceptual imagination its mediated objectual character.

"Modal imagination" is used here as a label for a certain sort of familiar mental act, and like other such categories, it resists straightforward definition. But its phenomenology is familiar. One has a positive

intuition of a certain configuration within a world, and takes that configuration to satisfy a certain description. When one modally imagines H₂O molecules, for example, one imagines a configuration of particles. To modally imagine Germany winning the Second World War, one might imagine a world in which certain German armies win certain battles and go on to overwhelm Allied forces within Europe. When one reflects on these imagined (parts of) worlds, they reveal themselves as (parts of) worlds in which there are H₂O molecules, or in which Germany won the Second World War.

Just as modally imagining that S goes beyond entertaining the proposition that S, modally imagining a world that verifies S goes beyond entertaining a proposition (even a highly specific proposition) that implies S. If this were all there were to modal imagination, then we could modally imagine any proposition trivially: just take the proposition itself, and conjoin with further propositions if necessary. But there are many propositions that we cannot easily modally imagine: complex unknown mathematical propositions M, for example. In these cases, we have no intuition of a world verifying M, even though we can entertain many specific propositions that imply M. So imagining a world is not merely entertaining a description. Of course it may be that imagining a world involves standing in *some* relation to a detailed description of that world (one presumably uses one's conceptual resources to imagine a world), but if so, this relationship goes beyond mere entertaining or supposing. Rather, it is a relation that is distinctive of modal imagination.

We can say that an imagined situation verifies S when reflection on the situation reveals it as a situation in which S. Understood this way, verification is a broadly epistemic relation, tied to certain rational processes. Importantly, verification is stronger than a mere evidential relation. We have seen that one can imagine situations in which no perceptual evidence is involved, as with the case of the nuclear force above. One can also imagine a situation in which one has strong evidence that S, such that the imagined situation is nevertheless epistemically compatible with $\sim S$: a situation where experimental results point to a certain sort of particle behavior, for example, or where usually reliable witnesses testify that someone committed a crime. In such cases, consideration of the imagined situation alone does not reveal that it is a situation in which S (as opposed to a situation in which there is strong evidence for S), so the imagined situations do not verify S. In this respect, verification of a statement by an imagined situation is broadly analogous to an entailment of one statement by another (a priori entailment, in the central cases): if it is coherent to suppose that the situation obtains without S being the case, then the situation does not verify S.

Just as imagining a unicorn does not entail the existence of the imagined unicorn, imagining a situation does not entail the existence of the imagined situation, and imagining a world does not entail the existence of the imagined world. Nothing here entails that one should be ontologically committed to situations or worlds at all. Rather, for our purposes these can be regarded as mere intentional objects, useful in characterizing the cognitive or phenomenological structure of modal imagination. It should also be noted that nothing here presupposes that when one imagines a situation or a world, there is a metaphysically possible situation or world that corresponds to the object of one's imagination. Again, these can simply be seen as *apparent* situations or worlds, of the sort represented in an act of imagination. For all that has been said so far, the imagination of situations and worlds may greatly outstrip the bounds of metaphysical possibility.

Indeed, it is arguable that one can modally imagine *S* when *S* involves an a priori contradiction. An example may be a case in which one imagines a geometric object with contradictory properties. In cases like this, one imagines a situation in something less than full detail. Another example may be a case when one imagines that a true mathematical claim (Goldbach's conjecture, perhaps) is false, by imagining a situation in which experts announce it to be false. In this sort of case, one might misinterpret the imagined situation as a situation in which *S*; here, the situation is merely one in which one has evidence for *S*.

To avoid cases like these, one can isolate a notion of *coherent modal imagination*. In this sense, *S* is positively conceivable when one can *coherently* modally imagine a situation that verifies *S*. A situation is coherently imagined when it is possible to fill in arbitrary details in the imagined situation such that no contradiction reveals itself. To coherently imagine a situation that verifies *S*, one must be able to coherently imagine a situation such that reasoning about the imagined situation reveals it as a situation that verifies *S*. This notion is our core notion of positive conceivability: I will henceforth say that *S* is positively conceivable when it is coherently modally imaginable.

One can then introduce *prima facie* and *ideal* versions of positive conceivability. *S* is *prima facie* positively conceivable when one can modally imagine a situation that one takes to be coherent and that one takes to verify *S*. *S* is ideally positively conceivable when *S* is *prima facie* positively conceivable and this positive conceivability cannot be undermined on idealized reflection. In effect, we can distinguish *prima facie* coherence from true coherence, and *prima facie* verification from true verification, where the "true" notions involve idealization on rational reflection. True coherence requires that arbitrary details can be filled in with no contradiction revealing itself on idealized reflection, whereas *prima facie* coherence requires merely the appearance of coherence. True verification requires that the imagined situation is revealed as a situation in which *S* even on idealized reflection, whereas *prima facie* verification requires merely the appearance that the imagined situation is a situation in which *S*. Then (invoking the "true" notions) one can say that *S* is ideally positively conceivable when one could coherently imagine a situation that verifies *S*.

When *S* is ideally positively conceivable, it must be possible in principle to flesh out any missing details of an imagined situation that verifies *S*, such that these details are imagined clearly and distinctly and such that no contradiction is revealed. It must also be the case that arbitrary rational reflection on the imagined situation will not undermine the interpretation of the imagined situation as one in which *S* is the case. These strictures are demanding, but they are not unreasonable. They are the strictures typically demanded of good thought-experiments, for example.

A typical philosophical thought-experiment starts with *prima facie* positive conceivability. A subject does not imagine a situation in fine detail: microphysical details are usually left unspecified, for example. Instead, a subject imagines a situation with certain important features specified, notes that a situation of this kind appears to verify *S*, and judges that the remaining details are not crucial: they can in principle be filled in to yield a full coherent conception of a situation that verifies *S*. For the thought-experiment to yield the intended conclusion, this *prima facie* judgment must be correct, so that *S* is ideally positively conceivable. If better reasoning would reveal that the details cannot be coherently filled in, or that the

situation does not truly verify S , then the thought-experiment will typically fail in its purpose. If the *prima facie* judgment is not defeatable in this way, however, then the thought-experiment succeeds, and S is ideally positively conceivable.

Clear cases of *prima facie* positive conceivability without ideal positive conceivability are surprisingly hard to come by. Possible examples might include the two cases above: imagining an impossible object, and imagining a situation in which mathematicians announce that M (for some false M). In these cases, however, even a moment's reflection is enough to undermine the positive conceivability. In the first case, one can easily detect a contradiction (or the inability to fill in crucial detail). In the second case, reflection reveals the situation as one in which one has evidence that M , but as not necessarily a situation in which M . So these cases will be *prima facie* positively conceivable under only the most superficial of reasoning processes.

A slightly better example of *prima facie* without ideal positive conceivability may be the Grim Reaper paradox (Benardete 1964; Hawthorne 2000). There are countably many grim reapers, one for every positive integer. Grim reaper 1 is disposed to kill you with a scythe at 1pm, if and only if you are still alive then (otherwise his scythe remains immobile throughout), taking 30 minutes about it. Grim reaper 2 is disposed to kill you with a scythe at 12:30 pm, if and only if you are still alive then, taking 15 minutes about it. Grim reaper 3 is disposed to kill you with a scythe at 12:15 pm, and so on. You are still alive just before 12pm, you can only die through the motion of a grim reaper's scythe, and once dead you stay dead. On the face of it, this situation seems conceivable - each reaper seems conceivable individually and intrinsically, and it seems reasonable to combine distinct individuals with distinct intrinsic properties into one situation. But a little reflection reveals that the situation as described is contradictory. I cannot survive to any moment past 12pm (a grim reaper would get me first), but I cannot be killed (for grim reaper n to kill me, I must have survived grim reaper $n+1$, which is impossible). So the description D of the situation is *prima facie* positively conceivable but not ideally positively conceivable.

Note that the mathematical case is a case in the subject has coherently imagined a situation but in which the imagined situation does not verify S on reflection, while the Grim Reaper and impossible object cases are cases in a situation has not been coherently imagined. Of course in both these cases, the problem is revealed by a little reflection. One might say that in this case (and in the mathematical case above), even if we have *prima facie* positive conceivability, we do not have *secunda facie* positive conceivability.

Cases of *secunda facie* positive conceivability without ideal positive conceivability seem to be extremely thin on the ground. Perhaps the best candidates involve rational but false beliefs in an *a priori* domain such as mathematics. In general, the details of an imagined situation will be irrelevant to the positive conceivability of a mathematical claim, since reflection suggests that the truth of the mathematical claim is independent of the imagined goings-on in the world. Rather, a mathematical claim will be positively conceivable insofar as there is rational reason to accept that claim; in that case, any imagined situation can be taken to verify the claim. One will have *secunda facie* positive conceivability without ideal positive conceivability when these reasons stand up to *secunda facie* scrutiny, but are undermined by ideal reflection. Frege's set of all sets may be such a case: Frege had good *a priori* reasons for accepting it that survived considerable reflection, but ideal (or at least Russellian) reflection reveals a deep contradiction.

If S is positively conceivable, S is negatively conceivable (in both the *prima facie* and ideal cases). If one can coherently imagine a situation verifying S, then one cannot rule out that S (though this interacts a little with the primary/secondary distinction below). The reverse is not the case, at least where *prima facie* conceivability is concerned: many statements are *prima facie* negatively conceivable without being *prima facie* positively conceivable. For example, as we saw above, many complex mathematical statements M are such that one cannot rule out M's truth, but one cannot imagine any situation (any part of a world) that verifies S. Something similar goes for statements in other *a priori* domains. And even in empirical domains, it may be that one cannot rule out M, but one cannot conceive of a situation in which M, due to limited powers of imagination, for example.

Clear cases of ideal negative conceivability without ideal positive conceivability are much harder to find. One might try mathematical statements that are true but not knowable *a priori* by any possible being. If there were such statements, they and their negations would be ideally negatively conceivable, but probably not ideally positively conceivable. But it is far from clear that there are any such statements. I will return to this matter later.

Positive conceivability, rather than negative conceivability, seems to be what most philosophers have had in mind when discussing conceivability. It is positive conceivability that corresponds to the sort of clear and distinct modal intuition that Descartes had in mind, and which reflects the practice in the method of conceivability as used in contemporary philosophical thought-experiments. When Yablo (1993) dismisses the first Goldbach example as not really being an instance of conceivability, he is in effect saying that negative conceivability is not true conceivability, and there is something to this.

Still, it must be conceded that negative conceivability is at least better defined than positive conceivability. The characterization of positive conceivability that I have given here, invoking the notion of a modally imagining a situation, cannot be considered a reductive definition. At best, it is something of a clarification. Nevertheless, there seems to be a reasonably clear intuitive notion in the vicinity, which most people seem to have a grasp on. It may be that the notion can be given a more rigorous definition, or it may be that it should be taken as primitive; this is one of the central open questions in the area.

The distinction between positive and negative conceivability bears at least some relation to van Cleve's (1983) distinction between strong and weak conceivability. According to van Cleve, P is strongly conceivable for a subject when the subject sees that P is possible; and P is weakly conceivable when the subject does not see that P is impossible. There is an obvious link between one reading of "seeing that S is impossible" and the idea of ruling out the hypothesis that S. And the notion of "seeing that S is possible" can be read as a sort of modal intuition that S of the sort that goes along with modally imagining that S.

I think that it is best not to import the notion of possibility so directly into a definition of conceivability, to avoid the threat of trivializing the link with possibility. In particular, there is a threat that the idealized version of seeing that S is possible will collapse into correctly judging that S is possible, which will be linked trivially to possibility. Still, the idea is closely related to the idea of coherently imagining a world (or a part of a world) that verifies S: both involve a sort of modal appearance. The main advantage of the

construal I have given is that it builds in no presupposition that the imagined world is metaphysically possible, or even that it seems metaphysically possible. It builds in some broadly modal elements, in the ideas of imagining a world, of coherence, and of verification. But importantly, the modalities here are cognitive or epistemic, and presuppose no tie to the metaphysical. To imagine a world is simply to engage in a distinctive and familiar sort of mental act; and the notions of coherence and verification are wholly grounded in rational notions. So there is no danger of trivializing the link between positive conceivability and possibility.

PRIMARY vs. SECONDARY CONCEIVABILITY. This distinction draws its motivation from Kripke's discussion of the necessary a posteriori. In the wake of Kripke's arguments that a posteriori statements such as "Hesperus is Phosphorus" are necessary, it has become a familiar observation that there is a sense in which "Hesperus is not Phosphorus" is conceivable, and a sense in which it is not. The first of these senses corresponds to primary conceivability; the second to secondary conceivability.

We can say that *S* is *primarily conceivable* (or *epistemically conceivable*) when it is conceivable that *S* is *actually* the case. We can say that *S* is *secondarily conceivable* (or *subjunctively conceivable*) when *S* conceivably *might have been* the case. This corresponds to two different ways of thinking about hypothetical possibilities: epistemically, as ways the world might actually be, and subjunctively, as counterfactual ways the world might have been. I have written more on these distinctions elsewhere, but I will give a short characterization here.

It is simplest to start with the case of positive conceivability. When one imagines a situation, one can consider it *as actual* (as a way the world might actually be), or one can consider it *as counterfactual* (as a way the world might have been). It is often the case that one will describe a situation differently depending on whether one considers it as actual or as counterfactual. We can say that *S* is primarily positively conceivable when one can coherently imagine a situation that verifies *S* when considered as actual, and that *S* is secondarily positively conceivable when one can coherently imagine a situation that verifies *S* when considered as counterfactual.

Primary conceivability is grounded in the idea that for all we know a priori, there are many ways the world might be. The oceans might contain H₂O or they might contain XYZ; the evening star and the morning star might be the same or distinct; and so on. We can think of these ways the world might be as *epistemic possibilities*, in a broad sense according to which it is epistemically possible that *S* if the hypothesis that *S* is not ruled out a priori. When *S* is epistemically possible, there are usually a number of imaginable situations such that if they actually obtain *S* will be the case. These situations can be taken to verify *S*, when they are considered as actual.

For example, it is epistemically possible in this sense that Hesperus is not Phosphorus (it is not a priori that Hesperus is Phosphorus). In the background of this epistemic possibility are many specific epistemically possible situations in which the heavenly bodies visible in the morning and evening are distinct. Upon consideration, these epistemically possible situations are revealed as instances of the epistemic possibility that Hesperus is not Phosphorus. There is a clear sense in which these situations *verify* the claim that Hesperus is not Phosphorus: for example, if one hypothetically accepts that such a

situation actually obtains, one should rationally conclude that Hesperus is not Phosphorus. This sort of relation among epistemic possibilities plays a central role in our thought.

When we consider situations as actual, we consider and evaluate them in the way that we consider and evaluate epistemic possibilities. That is, we say to ourselves: what if the actual world is really that way? One hypothetically assumes that the situation in question is actual, and considers whether, from that assumption, it follows that S is the case. If so, then the situation verifies S, when considered as actual. In the case above, for example, the situations in question (considered as actual) verify "Hesperus is not Phosphorus". So "Hesperus is not Phosphorus" is primarily positively conceivable.

(Primary conceivability is related to what Yablo (1993) calls "conceivability_{ep}", which requires that one can imagine believing something true with one's actual P-thought, but it is not quite the same. One difference is that primary conceivability does not require that a conceived situation contain a P-thought. So it is primarily conceivable that nothing exists, or that no-one thinks - these are not ruled out a priori, and are verified by certain situations considered as actual - but they are not conceivable in Yablo's sense.)

Negative, positive, prima facie, and ideal versions of primary conceivability are easy to formulate. We can say that S is primarily negatively conceivable when it is not ruled out a priori that S is actually the case, or more briefly, if S is not ruled out a priori. Positive primary conceivability, by contrast, requires coherently imagining a situation (considered as actual) that verifies S. Prima facie and ideal versions of these notions can be straightforwardly formulated as in the previous section. Primary positive conceivability implies primary negative conceivability, for both the prima facie and ideal versions, but the reverse is not obviously the case.

Primary conceivability is always an a priori matter. We consider specific ways the world might be, in such a way that the true character of the actual world is irrelevant. We can think of epistemic possibilities being entertained in such a way that empirical knowledge is suspended, and the reasoning that goes into evaluating them is a priori reasoning.

Secondary conceivability works quite differently. It is grounded in the idea that we can conceive of many counterfactual ways the world might have been, but is not. When we consider imagined situations as counterfactual, we consider and evaluate them in the way that we consider and evaluate counterfactual possibilities in the subjunctive mode. That is, we acknowledge that the character of the actual world is fixed, and say to ourselves: if the situation *had* obtained, what *would have been* the case? If we judge that had the situation obtained, S would have been the case, then we judge that the situation verifies S when considered as counterfactual.

Take an imagined situation in which the morning star is distinct from the evening star. Along with Kripke, we can say that if this situation had obtained, it would not have been the case that Hesperus was not Phosphorus. So when this situation is considered as counterfactual, it is revealed not as a situation in which Hesperus is not Phosphorus, but rather as a situation in which at least one of the objects is distinct from both Hesperus and Phosphorus (at least if we take for granted the actual-world knowledge that Hesperus is Phosphorus, and if we accept Kripke's intuitions). The reason is that (if Kripke is right) the

application of a term like "Hesperus" to a counterfactual situation depends on whether the actual Hesperus (i.e. the planet Venus) is present within that situation, and of course the actual Hesperus and the actual Phosphorus are one and the same. So when considered as counterfactual, this conceivable situation does not verify "Hesperus is not Phosphorus". More generally (if Kripke is right), there is no coherently imaginable situation, considered as counterfactual, that verifies "Hesperus is not Phosphorus". If so, "Hesperus is not Phosphorus" is not secondarily positively conceivable.

Unlike primary conceivability, secondary conceivability is often a posteriori. It is not secondarily conceivable that Hesperus is not Phosphorus, but one could not know that a priori. To know this, one needs the empirical information that Hesperus is actually Phosphorus. This aposteriority is grounded in the fact that the application of our words to subjunctive counterfactual situations often depends on their reference in the actual world, and the latter cannot usually be known a priori.

There are various ways to formulate *prima facie* and ideal versions of secondary conceivability. One might say that a subject *prima facie* secondarily conceives of S when the subject imagines a situation and judges that if that situation had obtained, S would have been the case. One can say that S is ideally secondarily conceivable if S is *prima facie* secondarily conceivable and if the secondary conceivability is not defeatable by idealized rational reflection and complete empirical knowledge. To avoid trivializing a link between conceivability and possibility here, it is probably best to restrict the empirical knowledge in question to nonmodal knowledge.

This characterizes positive versions of secondary conceivability. One might say that S is negatively secondarily conceivable when a priori reflection and empirical nonmodal knowledge reveals no incoherence in the hypothesis that S might have been the case. In any case, as secondary conceivability turns on a posteriori considerations, it will not be our central concern, and most of these varieties can be set to one side.

GAPS BETWEEN CONCEIVABILITY AND POSSIBILITY

With the distinctions above in play, it is relatively easy to classify potential gaps between conceivability and possibility.

(1) *Prima facie* conceivability is an imperfect guide to possibility.

Given that there is a gap between *prima facie* and ideal conceivability, it is only to be expected that there is a gap between *prima facie* conceivability and possibility. *Prima facie* conceivability judgments are sometimes undermined by continued rational reflection, isolating a contradiction or a misdescription in an apparently conceivable state of affairs. When this happens, then any grounds that the conceivability judgment provided for a claim of possibility will also be undermined.

This gap is widest in the case of *prima facie* negative conceivability judgments. When such a judgment is not backed by a corresponding *prima facie* positive conceivability judgment, it provides at best weak

evidence for possibility. Mathematical cases, such as the *prima facie* negative conceivability of both Goldbach's conjecture and its negation, provide an obvious source of gaps here. So likewise does any domain in which one might expect to find deep *a priori* truths.

Prima facie positive conceivability is a much better guide to possibility, but it is still imperfect. The case where one conceives of mathematicians announcing a proof of Goldbach's conjecture (or its negation) is best seen as a case where a superficial *prima facie* positive conceivability judgment is undermined by a moment's reflection. Other cases of *prima facie* conceivability without possibility may be provided by the Grim Reaper paradox and the case of impossible objects.

Cases of *secunda facie* positive conceivability, where a *prima facie* positive conceivability judgment survives a reasonably searching process of rational reflection, are a still stronger guide to possibility. In the great majority of cases with a gap between *prima facie* and ideal positive conceivability, the *prima facie* judgment is easily undermined by a little reflection. Gaps between *secunda facie* positive conceivability and ideal positive conceivability seem to be very rare, although perhaps the Frege case is an example.

In any case, if we are looking for a notion of conceivability such that conceivability tracks possibility perfectly, we must focus on ideal conceivability. In this sense conceivability is not a merely psychological notion; it is a *rational* notion, in much the same way that a priority and rational entailment are rational notions. If there is to be a plausible epistemic/modal bridge, it will be a bridge between the rational and modal domains.

(2) Primary conceivability is an imperfect guide to secondary possibility.

The other standard source of gaps between conceivability and possibility arises from Kripkean cases. It is often said that it is conceivable that Hesperus is not Phosphorus, or that water is not H₂O, or that heat is not the motion of molecules, but none of these states of affairs are in fact possible. In these cases we have *a posteriori* necessities and impossibilities, out of the reach of *a priori* methods.

There are a couple of things to be said here. Clearly, the main sense in which these states of affairs are conceivable involves primary conceivability. As discussed earlier, the states of affairs in question are not secondarily conceivable. At best, they might be *prima facie* secondarily conceivable for a subject lacking relevant empirical knowledge. They will not be *prima facie* secondarily conceivable for a subject with the relevant knowledge, and they will not be ideally secondarily conceivable as that notion is spelled out above.

One might then try to save a conceivability/possibility link by suggesting that ideal secondary conceivability entails possibility. This thesis is not implausible, but it is not helpful for our purposes here. The reason is that secondary conceivability, and especially ideal secondary conceivability, is deeply *a posteriori*. So even if secondary conceivability is a guide to possibility, it will yield no *a priori* access to modality.

(Around this point, it seems to me that the otherwise excellent discussions of conceivability and possibility by Menzies (1998), van Cleve (1983), and Yablo (1993) all give up too soon, settling for conceivability/possibility theses that are more attenuated than necessary.)

If we are interested in modal rationalism, we should instead focus on ways in which primary conceivability might still be a guide to possibility. Even if it is conceded that strictly speaking, it is not possible that water is H₂O, it can still be argued that the primary conceivability of "water is not H₂O" is revealing something about metaphysical possibility. When we apparently conceive of a world in which water is not H₂O, we conceive of a situation in which some other substance (XYZ, say) is the clear liquid surrounding us in the oceans and lakes, and so on. And this situation is indeed metaphysically possible - so our act of conceiving has indeed yielded access to a possible world. It is just that in a certain sense we have misdescribed it in calling it a world where water is not H₂O, or a world in which water is XYZ. If Kripke is right, it is in fact a world in which XYZ is watery stuff but not water, and a world in which the only water that exists is H₂O.

Further: there remains a sense in which a world with XYZ in the oceans can be seen as satisfying the statement "water is not H₂O". Here, I will give a very brief version of a story that I have told in more detail elsewhere (e.g. Chalmers 1996, forthcoming b; see also Evans 1977, Davies and Humberstone 1980, and Jackson 1998).

As discussed earlier, there is clearly a broad sense in which it is *epistemically* possible that water is not H₂O, in that the hypothesis is not ruled out a priori. Intuitively, there are ways our world could turn out such that if they turn out that way, it will turn out that water is not H₂O. And if we consider the XYZ-world as an epistemic possibility - that is, we consider the hypothesis that the world with XYZ in the oceans is *our* world - then this epistemic possibility can be seen as an instance of the epistemic possibility that water is not H₂O. We can rationally say "if our world turns out to have XYZ in the oceans (etc.), it will turn out that water is not H₂O". This might be put as a simple indicative conditional: "if XYZ is in the oceans and lakes (etc.), then water is XYZ". Compare: "if Prince Albert Victor committed those murders, then he is Jack the Ripper". Here, the indicative conditional "if P, then Q" can be evaluated using the Ramsey test: if one hypothetically accepts the belief that P, does one arrive at the conclusion that Q?

All this reflects the fact that we have a systematic way of evaluating and describing epistemic possibilities that differs from our way of evaluating and describing subjunctive counterfactual possibilities. In both cases, we consider and describe worlds, but in the epistemic case, we consider them as actual, whereas in the subjunctive case, we consider them as counterfactual. And these two modes of consideration of a world yield two ways in which a world might be seen to satisfy a sentence. When the XYZ-world is considered as actual, it satisfies "water is XYZ"; when it is considered as counterfactual, it does not.

Given a statement S and a world W, the *primary intension* (or *epistemic intension*) of S returns the truth-value of S in W considered as actual. Three heuristics for evaluating the primary intension of S in W correspond to the three tests mentioned above. One can appeal to direct evaluation of epistemic possibilities: is the epistemic possibility that W is actual an instance of the epistemic possibility that S? One can appeal to indicative conditionals (evaluated by the Ramsey test): if W is the case, is S the case?"

Or one can appeal to the "turns out" locution: if W turns out to be actual, will it turn out that S?

Primary intensions can be formally defined in terms of a priori entailments. In particular, we can say that the primary intension of S is true in W if the material conditional "if W is actual, then S" is a priori: that is, if the hypothesis that W is actual and S is not the case can be ruled out a priori. S's primary intension is false in W if the conditional "if W is actual, then $\sim S$ " is a priori; and S's primary intension is indeterminate at W if neither of these conditionals is a priori. For example, the hypothesis that the XYZ-world is actual and water is H₂O can plausibly be ruled out conclusively by rational reflection alone. If so, the material conditional "if the XYZ-world is actual, then water is not H₂O" is a priori, and the primary intension of "water is H₂O" is false in the XYZ-world. For more on the definition of primary intensions, see the further discussion below.

Primary intensions are grounded in the *epistemic* evaluation of statements in worlds: that is, the evaluation of statements in worlds considered as actual. One can also define the notion of a *secondary* (or *subjunctive*) intension, grounded in the subjunctive evaluation of statements in worlds: that is, the evaluation of statements in worlds considered as counterfactual. The secondary intension of a statement S is the function that maps a world W to the truth-value of S in W considered as counterfactual. These correspond to a much more familiar notion of intension in contemporary philosophy, so I will say less about them here.

To characterize secondary intensions with a heuristic, one can appeal to subjunctive conditionals: if W had obtained, would S have been the case? Or one can appeal directly to intuitions about counterfactual possibilities: is W a counterfactual possibility in which S would have been the case? Heuristics of this sort are frequently invoked by Kripke in his evaluation of possible worlds; and his corresponding influential claims about possibility are almost always grounded in subjunctive claims about what might have been the case. So the intensional notions that arise from Kripke's work are all closely tied to secondary intensions.

A paradigmatic example involves the subjunctive evaluation of a statement such as "water is XYZ" at the XYZ-world, a world that is similar to our own except that the watery liquid in the oceans and lakes is XYZ. If Kripke and Putnam are correct, then if the watery stuff in the oceans and lakes had been XYZ, it would nevertheless not have been the case that water was XYZ: at best, XYZ would have been watery. Corresponding, W does not seem to represent a counterfactual possibility in which water is XYZ. So the secondary intension of "water is XYZ" is false at the XYZ-world.

We can then say that S is *primarily possible* (or 1-possible) if its primary intension is true in some possible world (i.e. if S is true in some world considered as actual). S is *secondarily possible* (or 2-possible) if its secondary intension is true in some possible world (i.e. if S is true in some world considered as counterfactual). Primary and secondary necessity can be defined analogously.

Secondary possibility and necessity correspond to the standard conception of what it is for a statement to be metaphysically possible or necessary. For example, "water is H₂O" is plausibly 2-necessary, and "water is XYZ" 2-impossible, reflecting their metaphysical necessity and impossibility (as standardly

understood) of respectively. On this understanding, we can say that a statement is metaphysically necessary iff it has a necessary secondary intension.

Primary possibility and necessity correspond much more closely to epistemic notions such as apriority. It is clear that when *S* is a priori, it will have a necessary primary intension, so it will be 1-necessary. Whether the reverse entailment (from 1-necessity to apriority) holds is one of the central issues in this paper, but for now we can note that at least the clearest cases of 1-necessary statements are all plausibly a priori: witness " $2+2=4$ ", or "Hesperus, if it exists, is visible in the evening" (1-necessary and a priori), as opposed to "tables exist" and "water is H₂O" (1-contingent, and a posteriori).

The existence of primary and secondary intensions suggests that expression tokens have a complex semantic value that involves both intensions. These intensions will play important roles when the expression is embedded in different contexts. In constructions such as "it might have been the case that *S*" and subjunctive conditionals, *S*'s secondary intension will be relevant. In constructions such as "it is a priori that *S*" and indicative conditionals, *S*'s primary intension will be relevant. Both intensions are part of the content of *S* in both contexts: it is just that the different contexts exploit different aspects of *S*'s content. The propositional content of *S* might be understood in a number of different ways, but if one holds that the apriority and necessity of *S* is a function of the proposition that *S* expresses, then the proposition expressed by *S* will be reducible to neither its primary intension nor its secondary intension, but will rather be something that involves at least the structure of both.

We can now see how primary conceivability can act as a guide to possibility. When we find it conceivable that water is not H₂O, there is no possible world that satisfies "water is not H₂O" when the world is considered as *counterfactual*, but there is a possible world that satisfies "water is not H₂O" when the world is considered as *actual*. Put differently, the secondary intension of "water is H₂O" is true in no world, but the primary intension is true in some (centered) worlds. The XYZ-world, and other centered worlds that we might conceive of when we conceive that water is not H₂O, all satisfy the primary intension of "water is H₂O".

We can put this by saying that primary conceivability is an imperfect guide to secondary possibility, but is a much better guide to primary possibility. In all the Kripkean cases in which *S* is primarily conceivable, *S* is also primarily possible (or at least Kripke's discussion gives no reason to deny this). There is a (centered) possible world satisfying the primary intension of "Hesperus is not Phosphorus" (e.g. a world where heavenly bodies visible from the center in the morning and the evening are distinct), of "heat is not the motion of molecules" (e.g. a world where something else causes heat sensations), and so on. These worlds are all first-class metaphysical possibilities.

So Kripke's examples are entirely compatible with the thesis that conceivability is a guide to possibility. We just need to make sure that the relevant notions are aligned: primary conceivability is a guide to primary possibility, and secondary conceivability is a guide to secondary possibility. This is no surprise: it would be odd to expect conceivability of a situation considered as actual to be a guide to possibility of a world considered as counterfactual, or vice versa! So we are still left with significant a priori access to the space of possible worlds.

Sideline: On Defining Primary Intensions

Primary intensions are intensions that capture the distinctive way a statement is used to describe and evaluate epistemic possibilities. The primary intension of a statement could be defined in various ways, but the most useful definition is that in terms of a priori entailments: the primary intension of S is true at W if the material conditional "if W is actual, then S" is a priori. I elaborate and defend this conception of a primary intension in other work; here I will make a few observations about the definition of primary intensions and about their properties. This material can be skipped by those who are not interested in the fine details of the two-dimensional framework.

(i) For a world to be considered as actual, it must be a *centered* world: a world marked with a specified individual and time. The reason is that an epistemic possibility is not completely determined until one's "viewpoint" is specified. For example, an objective description of the world will not allow me to settle the question of whether I am in Australia or in the US, but a "centered" specification will do this. The hypothesis that a centered world W is actual, for me, will include the hypothesis that I am the being marked at the center and that now is the time marked at the center. A primary intension can then be seen as a function from centered world to truth-values. The primary intension of "I am a philosopher", for example, will be true at those centered worlds in which the subject at the center is a philosopher.

(ii) The evaluation of a conditional involving "If W is actual..." requires a *canonical description* of W. We can say that the primary intension of S is true at W if the material conditional "if D, then S" is a priori, where D is a canonical description of W. The notion of a canonical description can be elaborated in various different ways, which are too complex to discuss in detail here. One needs to isolate a sort of neutral vocabulary in which worlds can be described, and to require a certain sort of complete description within this vocabulary. On the first point, a neutral expression might be seen intuitively as one which behaves the same way in epistemic and subjunctive evaluation (plus indexicals such as "I" and "now" to handle centering). On the second point, one might require a complete description to be ontologically complete, or qualitatively complete, or epistemically complete, in the terms from later in this paper. If the theses of this paper are correct, these different notions of completeness are coextensive. If the theses of this paper are incorrect, these notions may come apart, yielding different primary intensions. In that case, it is probably best to require epistemic completeness in the definition.

(iii) Primary intensions are defined here as functions over (centered) possible worlds. One can also define a closely related intension as a function over an independently characterized *epistemic space* of maximal epistemic possibilities. Epistemic space is not defined in terms of metaphysical possible worlds, but rather in terms of epistemic notions such as apriority: maximal epistemic possibilities correspond roughly to maximally specific a priori consistent hypotheses concerning the actual world. One can define an intension over this space much as one defines a primary intension. In other work (e.g. Chalmers forthcoming a,b), I have called this an *epistemic intension*.

What is the relationship between the two notions? This relationship turns on the relationship between epistemic space and the space of centered possible worlds, which in turn is closely tied to the relationship

between ideal negative conceivability and primary possibility. If this paper's theses are correct, there is a direct correspondence between the two spaces, so that primary intensions as defined here and epistemic intensions are almost identical. If this paper's theses are incorrect, then the definitions come apart: there will be maximal epistemic possibilities that correspond to no centered possible worlds, so the intensions will be defined over different spaces.

For many purposes, especially within the epistemic domain, the notion of an epistemic intension is more fundamental. For example, necessity of epistemic intension is constitutively tied to apriority and other epistemic notions, independently of any views about metaphysical possibility. So epistemic intensions can be used for epistemic purposes regardless of one's further views. For present purposes, however, the link between the epistemic and metaphysical domain is the central focus, so I focus here on primary intensions understood as functions over metaphysically possible worlds. If what I say in this paper is correct, the two intensions ultimately collapse into one.

(iv) The primary intension of some terms can vary between speakers. For example, Leverrier might use "Neptune" to pick out whatever causes certain orbital perturbations within a world, whereas an acquaintance might use it to pick out (roughly) whatever Leverrier refers to with the name, irrespective of any perturbing role. If so, their primary intensions will vary accordingly. All this reflects the fact that certain conditionals of the form "if W is actual, then Neptune is such-and-such" are a priori for Leverrier but not for his acquaintance. This happens not because of any difference in their rational capacities (which we are idealizing away from), but because of differences in the inferential roles associated with the term. Something similar can happen with most names and natural kind terms.

It follows that primary intensions are not candidates for linguistic meaning, the sort of meaning common to all tokens of an expression type, at least where names and natural kind terms are concerned. (See Chalmers forthcoming a.) To accommodate this phenomenon, primary intensions should be associated in the first instance with expression tokens (or perhaps with types as used on occasions), not with expression types. We can define primary intensions more precisely by saying that the primary intension of a statement token S (used by a speaker) is true in W if the material conditional "if W, then S" is a priori for the speaker. Here, a sentence T will be a priori for a speaker if the belief (or the hypothesis) that the speaker expresses with T could be conclusively justified, on ideal rational reflection, with justification independent of experience. On this account, different material conditionals will be a priori for Leverrier and his wife, so their primary intensions for "Neptune" will differ accordingly.

Note that the notion of apriority (whether speaker-relative or speaker-independent) requires the same sort of rational idealization as that present in the notion of ideal conceivability. I have defended the claim that relevant conditionals are a priori elsewhere (see also the discussion of scrutability later in this paper). If someone is skeptical about this, or skeptical about the very notion of apriority, it may nevertheless remain plausible that the material conditionals in question have *some* distinctive epistemic status which can be used to define a corresponding notion of primary intension.

(v) To evaluate the primary intension of S in W, it is not required that W contain a token of S. The heuristics and definition above give no special role for such a token, even when it is present. One could

define the *contextual intension* of S as a function defined across worlds containing a token of S at the center, returning the truth-value of that token. This would be a useful notion, but it would be quite different from the current notion, which is grounded in the epistemic domain. Contextual intensions turn on the context-dependence of a statement's semantic value, while primary intensions turn on the use of a statement in evaluating epistemic possibilities. Contextual intensions are more closely related to Kaplan's notion of "character" and Stalnaker's notion of a "diagonal proposition" than the current notion of a primary intension.

To see some differences, note that the contextual intension of statements such as "language exists" will plausibly be nowhere false, but the primary intension of "language exists" will be false in many (language-free) centered worlds. This reflects the (broad) epistemic possibility of such worlds: it is not a priori that language exists. Something similar applies to "nothing exists" (whose primary intension is true of an empty world) and many claims about thinkers and about language. The contextual intension also requires an account of what it takes for a token to count as an instance of S's type, raising problems (pointed out by Block and Stalnaker 1999), that tend to break the link between contextual intensions and epistemic notions. If we individuate S's type orthographically, "bachelors are unmarried" has a contingent contextual intension; if we individuate by familiar sorts of semantic content, "water is XYZ" has a necessarily false contextual intension; if we individuate by "narrow content" or some such, then we need an independent account of that sort of content. This issue does not arise for primary intensions. The effect is that primary intensions are much more directly connected to the epistemic domain than are contextual intensions.

(vi) Yablo (this volume) considers various ways in which the epistemic evaluation of statements in worlds might be defined. He rejects both the indicative conditional heuristic and the "turns out" heuristic, on the grounds that they give the wrong result in certain metalinguistic cases. For example, 'tail are wings' should be false in a world (considered as actual) where 'tail' is used to refer to wings. But Yablo suggests that the indicative conditional "if 'tail' refers to wings, then tails are wings" is intuitively correct, as is "if it turns out that 'tail' refers to wings, then it will turn out that tails are wings".

I think these judgments of intuitive correctness are unclear, and I think there is at least a reasonable reading of the locutions on which the conditionals in questions are incorrect. (Compare the reasonable: "if 'tail' refers to wings, 'tail' does not refer to tails"). But even if Yablo were right about this, it would show only that the heuristics are imperfect, giving the wrong results in special cases. The problem does not arise for the definition I have given here, in terms of a priori material conditionals.

To see this, note that the claim that 'tail' refers to tails is not a priori, but represents substantive a posteriori metalinguistic knowledge. It is a posteriori that the orthographic string 'tail' means anything at all, and it is a posteriori that it means what it does. So there is no a priori entailment from claims involving "'tail'" to corresponding claims involving 'tail'. And in particular, there is no a priori entailment from "'tail' refers to wings" to 'tails are wings'. More generally, there are plausibly no substantive a priori connections between claims about the orthographic string 'tail' and claims about tails, since any inferential connections between these claims rest on a posteriori metalinguistic knowledge. If so, the way that 'tail' is used in such a world will be irrelevant in evaluating the primary intension of statements such as 'tails are wings' in that world.

In particular, there is no danger that 'tails are wings' will be true in a world (considered as actual) in which 'tail' refers to wings.

Someone might suggest that there is a semantic concept of "'tail'" that builds in semantic constraints as well as orthographic constraints, so that it is a priori that 'tail' refers to tails. But then the worlds Yablo considers, in which the orthographic string 'tail' refers to wings, will not be worlds in which 'tail' (construed semantically) refers to wings, so there is no danger that they will be worlds (considered as actual) in which tails are wings. Either way, the usage of the orthographic string 'tail' in a world will be irrelevant in evaluating the primary intension of 'tails are wings' in that world.

Yablo himself endorses a mixed view, on which it is a priori that 'sister' refers to sisters (like the semantic view), but on which it is not a priori that 'sister' refers to female siblings (like the orthographic view), even though it is a priori that sisters are female siblings. A mixed view like this cannot be accommodated in the current framework: the framework requires that apriority is preserved under a priori entailment, but Yablo's view violates this. (For the relevant A, B, C, it is a priori that A, that B, that if A and B then C, but not that C). But it can plausibly be argued that this violation of closure is reason enough to reject Yablo's mixed view. (The view seems to be grounded in an idiosyncratic conception of a priori knowledge, on which a priori knowledge that S depends on metalinguistic knowledge concerning 'S'. I think that such a view should clearly be rejected.) In any case, it seems that any residual problems here arise from Yablo's somewhat idiosyncratic view of these metalinguistic cases, and not from the cases themselves.

For his own view of the epistemic evaluation of statements in worlds, Yablo endorses a "could have turned out" heuristic: "If it had turned out that W, would it have turned out that S?". Although I have occasionally used this heuristic myself in earlier work, I am less comfortable with it than with the "turns out" heuristic, as the subjunctive conditional here can easily be read non-epistemically, and it is too close to the subjunctive "If it had been that W, it would have been that S" for comfort. (I am also worried that Yablo's paper has the wrong title - most uses of "coulda", "woulda", and "shoulda" go with secondary intensions (!), as characterized below.) Still, it may be that there is at least a reading of this locution that gives approximately correct results in most cases. (End of sideline.)

(3) *Positive conceivability is a better guide to possibility than negative conceivability.*

We have already seen that prima facie negative conceivability is a relatively weak guide to possibility. The canonical case here is the prima facie negative conceivability of both Goldbach's conjecture and its negation. It should be noted that neither of these cases is backed by a corresponding prima facie positive conceivability. (There is a "positive" case in which mathematicians announce the proof, but this yields at best a very tenuous prima facie positive conceivability judgment, and raises separate issues).

So at least where prima facie conceivability is concerned, positive conceivability is a much better guide to possibility than negative conceivability. This fits the usual practice in philosophy, where the conceivability judgments that are usually taken as evidence of possibility are almost always positive conceivability judgments. (For just this reason, the Goldbach case was never a very compelling

counterexample to this practice.)

With ideal conceivability, things are less clear. Certainly ideal positive conceivability is at least as good a guide to possibility as ideal negative conceivability, since the former entails the latter. What is less clear is whether there are cases of the latter without the former, and if so, whether those cases correspond to possibilities.

The most obvious potential case here is an extension of the Goldbach case above. If either Goldbach's conjecture or its negation is provable, or otherwise knowable a priori, then only one will be ideally negatively conceivable. But perhaps (as noted earlier) there are some true or false mathematical statements whose truth-value cannot be settled even by ideal rational reflection? If so, we would have cases of ideal negative conceivability without ideal positive conceivability and without possibility. It is not at all clear that cases of this type exist, however. I will discuss this and other potential counterexamples to a link between ideal negative conceivability and possibility later. It seems that there are at least no clear counterexamples, so a link between ideal negative conceivability and possibility remains tenable.

Overall, we can say that both ideal positive conceivability and ideal negative conceivability are promising as guides to possibility, but that the former is in a slightly better position to be a perfect guide than the latter, due to its added strength.

CONCEIVABILITY/POSSIBILITY THESES

To summarize: if any variety of a priori conceivability entails possibility, it must be a variety of ideal primary conceivability, and the variety of possibility that is entailed must be primary possibility. And positive conceivability is always at least as good a guide to possibility as negative conceivability. This leaves us with the following as the most plausible entailment between conceivability and possibility:

(1) *Ideal primary positive conceivability entails primary possibility.*

That is, if S is ideally primarily positively conceivable, then there is some metaphysically possible centered world satisfying S's primary intension (or that satisfies S when considered as actual).

We have also left open the status of the following:

(2) *Ideal primary negative conceivability entails primary possibility.*

For completeness, I note that the following two theses also remain plausible, although neither suffices for a thoroughgoing modal rationalism.

(3) *Secunda facie primary positive conceivability is an extremely good guide to primary possibility.*

(4) *Ideal secondary (positive/negative) conceivability entails secondary possibility.*

We have seen that the first thesis is compatible with the standard clear counterexamples to a link between conceivability and possibility, as is the second (although there are some unclear counterexamples that may threaten the second). So if there are any counterexamples to these two theses, they must come from a different source, and their existence will gain no support from the standard cases.

For my part, I think that thesis (1) is almost certainly true, and that thesis (2) is very likely true. In most of the rest of the paper, I will discuss what counterexamples to these theses would involve, and give a quick sketch of reasons to think the theses true.

Note that because the conceivability and possibility of a statement is speaker relative, the conceivability-possibility theses above must be interpreted in a speaker-relative way: if S is conceivable for a speaker, S is possible for that speaker. There are two main sources of speaker-relativity here: variation in cognitive capacity, which affects *prima facie* conceivability, and variation in primary and secondary intensions of terms, which affects primary conceivability/possibility and secondary conceivability/possibility respectively.

For the central theses (1) and (2), only the second sort of variation is relevant. This variation manifests itself in phenomena such as the following: "Neptune does not perturb the orbit of Uranus" may be ideally primarily conceivable for Leverrier's acquaintances but not for Leverrier himself; it will also be primarily possible for Leverrier's acquaintances, but not for Leverrier himself. In a similar way, "I am not David Chalmers" may be ideally secondarily conceivable for you but not for me; it is also secondarily possible for you but not for me. Because the variation here affects conceivability and possibility equally, it does not threaten an inference from conceivability to possibility. It suggests at most that in cases where this variation is present, the inference must be speaker-relative.

One might worry that because the notion of ideal conceivability itself involves the notion of possibility (for example, in claims about what some possible being could conceive, or about what is defeatable), there is a danger of circularity. There are a few different issues here. First, one might worry that this rules out a reduction of possibility to conceivability. But I am not trying to give such a reduction, but am simply investigating the connection between the two notions. Second, one might worry that circularity will make the conceivability-possibility thesis trivial. But the notion of possibility enters into the definition of conceivability in such a roundabout way that the thesis clearly remains substantive. Finally, one might worry that defining conceivability in terms of possibility renders conceivability toothless as an epistemic guide to possibility, and so defeats modal rationalism. But this is not so: modal rationalism holds that modality is *a priori accessible*, and so invokes the notion of possibility in a precisely parallel manner. If ideal conceivability tracks possibility, then modal facts are rationally accessible, as required.

(If one wanted to give a reductive account of possibility in terms of conceivability, there is one strategy that is worth trying. Instead of invoking possible beings in the definition of conceivability, one could invoke conceivable beings. There might then be a sort of bootstrapping definition. First, the notion of

conceivability would be grounded in our own prima facie conceivings. Second, we can conceive of beings who are better reasoners than us, with fewer cognitive limitations. Third, those beings could presumably conceive of better reasoners still. And so on. It is not out of the question that this process might lead to some sort of limit or fixed point. If so, one might obtain a recursive (not noncircular) definition of possibility in terms of conceivability. I cannot pretend that this matter is entirely clear, however.)

Does this account leave room for modal error? If theses (1) and (2) are correct, then modal errors arising from conceivability judgments will stem either from the difference between prima facie and ideal conceivability or from the difference between primary and secondary conceivability (and possibility). These modal errors will fall into one or more of the following classes:

- (i) Prima facie negative conceivability judgments can go wrong in cases where a "deep" a priori contradiction is not revealed by prima facie reasoning.
- (ii) Prima facie positive conceivability judgments can go wrong when (a) an imagined situation that is taken to verify P does not in fact verify S, upon rational reflection; or when (b) an imagined situation is not coherently imagined, because of the failure to notice a deep contradiction, or because of the inability to fill in crucial details.
- (iii) Primary conceivability judgments can go wrong if a subject mistakenly expects them to be a guide to secondary possibility.
- (iv) Prima facie secondary conceivability judgments can go wrong as a guide to secondary possibility when a subject is misinformed about relevant nonmodal empirical facts, and perhaps when an incautious subject is merely ignorant of those facts.

FROM NEGATIVE TO POSITIVE CONCEIVABILITY

In the remainder of the paper, I will focus on the status of the central conceivability-possibility theses (1) and (2), discussing what is required in order for them to be true, what form counterexamples must take, whether there are any plausible counterexamples, and what might ground the truth of the theses. Given space limitations, this discussion will only scratch the surface, but I hope to convey at least a broad view of the terrain. In discussing these theses, we can restrict our attention mostly to ideal primary positive conceivability and ideal primary negative conceivability, and to primary possibility. "Ideal primary" should be understood throughout, in references to positive and negative conceivability, and "possibility" should always be read as primary possibility. The speaker-relativity of the relevant claims should also be understood throughout.

I will first address the question of whether (ideal primary) negative conceivability entails (primary) possibility. Here I will factor out the question (addressed later) of whether positive conceivability entails possibility, and address the question of whether negative conceivability entails positive conceivability.

I call the (empty or nonempty) class of statements that are negatively conceivable but not positively conceivable the *twilight zone*. Potential members of the twilight zone come from two sources: inscrutabilities and open inconceivabilities.

INSCRUTABILITIES:

The class of inscrutabilities can be introduced by considering an attractive thesis about truth and reference.

SCRUTABILITY OF TRUTH AND REFERENCE: Once we know how the world is qualitatively, we are in a position to know what our terms refer to and whether our statements are true.

Take the case of reference first. Often we do not know what our terms refer to, but this knowledge is usually grounded in some qualitative ignorance of the way the world is. Given enough qualitative information (typically information about physical and mental states, although more on this later), we are in a position to know what our terms refer to. This reflects common practice in the theory of reference: in thinking about reference of a term in actual and hypothetical situations (considered as actual), it suffices to give a complete enough qualitative description of relevant features of those situations. From here, reference can be determined.

There are a few difficulties with the thesis of scrutability of reference. The first is that it is not entirely obvious what it means to "know what a term refers to". Presumably this is to be able to give some sort of alternative description of the referent; but just which alternative descriptions qualify? The second is that there may be a degree of indeterminacy in the reference of our terms over and above what is present in the truth-values of our statements. Examples might include terms like "number" and "symphony", or Quine's examples of mass-nouns and count-nouns in Japanese (see Benacerraf 1965, Horgan 1986, and Quine 1961 respectively). In these cases, it seems that there are multiple ways to assign referents to our terms, each of which capture our intuitions about the truth-values of our statements, insofar as these truth-values are determinate. In these cases it is not at all clear that there is a fact of the matter between these assignments of reference.

For both of these reasons it is easier to focus on the scrutability of truth. On the first issue, there is no analogous problem making sense of what it is to know the truth-value of a statement. On the second issue, almost all of the indeterminacies discussed above drop out when it comes to the truth-values of statements. (The exception may be such statements as "the number two is a set of sets", and the like, but now we are at least down to an isolated problem in the metaphysical domain, as opposed to a problem that arises with every use of the word "two".) And most of the intuitive backing behind the scrutability of reference (e.g. that given enough qualitative information, we can know who Jack the Ripper is) is reflected in the scrutability of truth (e.g. that given enough qualitative information, we can know whether Jack the Ripper was Prince Albert Victor).

The scrutability of truth can be formulated somewhat more precisely as follows:

SCRUTABILITY OF TRUTH (second pass): If D is a complete qualitative description of the world, then for all T, $T \rightarrow (D \text{ implies } T)$.

Here and elsewhere, ' $A \rightarrow B$ ' is a material conditional, and A implies B iff ' $A \rightarrow B$ ' is a priori. So the scrutability thesis says, in effect, that all truths are derivable through a priori reasoning from a complete qualitative description of the world. The thesis can also be weakened somewhat to hold that all truths are derivable from a complete *enough* qualitative description, thus avoiding the need to invoke a description of the whole world for every truth, but I will use the simpler if less practical formulation in what follows.

This way of putting things is not only more precise than "Once we know A, we're in a position to know B"; it also overcomes a problem posed by the paradox of knowability. Let P be a truth that I don't currently know, and let Q be "P and I don't know that P". Then Q is true but unknowable. (To know Q, I would have to know P; but once I know P, then Q is false). So Q is in danger of coming out inscrutable on the first formulation - it is a truth such that having full qualitative information about the world doesn't suffice to know it. But for all this it may remain the case that Q is implied by a full qualitative description of the world. The "paradox" gives no special reason to deny that given such a complete qualitative description D, I can know a priori that if D, then Q.

We can also put the scrutability thesis in terms of *epistemic completeness*, where an epistemically complete statement is one that, roughly speaking, epistemically settles everything that could be settled. More precisely, let us say as before that a statement D is epistemically possible (in the broad sense) when D is not ruled out a priori: i.e. when it is not a priori that $\sim D$ (or that $\sim \text{det}(D)$, to cover cases of indeterminacy), i.e., when D is ideally negatively conceivable). Then

A statement D is *epistemically complete* iff (i) D is epistemically possible, and (ii) for all F, if D&F is epistemically possible, then D implies F.

Then the scrutability thesis, reformulated, says that a complete qualitative description of the world is epistemically complete. The second formulation implies the first, since if D is a complete qualitative description of the world and T is the case, then D&T is true, so D&T is epistemically possible, so (by the second formulation) D implies T. In the other direction: if D&F is epistemically possible, then D does not imply $\sim F$, so (by the contrapositive of the first formulation) $\sim \sim F$, so F, so (by the first formulation) D implies F. (Worries about indeterminacy are handled by the observation that if D&F is epistemically possible, D does not imply $\sim \text{det}(F)$, so the indeterminacy of F is excluded.)

The residual unclarity, of course, is in the notion of a "complete qualitative description of the world"? What counts as a complete qualitative description? One idea is that a complete enough qualitative description is one that specifies all truths; but this will not do for our purposes, since it renders the scrutability thesis trivial. Intuitively, a qualitative description of the world is a "neutral" description from which many other truths might be derived.

A second and promising idea says that a complete qualitative description is a complete description in terms of fundamental natural properties (plus indexical information). That is, it involves a description in terms of fundamental microphysical properties (perhaps such as mass, charge, position, and spin), and perhaps also in terms of those fundamental properties (if any) that are not microphysical (on some nonmaterialist views, phenomenal or protophenomenal properties). So understood, the scrutability thesis would come to the claim that the fundamental natural truth about the world, in conjunction with indexical truths, implies (a priori) all truths.

We might formalize this by understanding this sort of description of the world as an *ontologically complete* description of the world: roughly speaking, one that metaphysically necessitates all truths about the world. In order a resulting scrutability thesis to be tenable, the relevant sort of metaphysical necessitation must be 1-necessitation (recalling that a statement is 1-necessary if its primary intension is true in all centered metaphysically possible worlds). More precisely, we can say:

A statement *D* is *ontologically complete* if (i) *D* is 1-possible, and (ii) if *D*&*F* is 1-possible, then *D*->*F* is 1-necessary.

The resulting scrutability thesis is:

STRONG SCRUTABILITY: If *D* is an ontologically complete truth, then *D* is epistemically complete.

The strong scrutability thesis is interesting and not obviously false, and is closely related to thesis (2) connecting ideal negative conceivability and possibility (in fact it is a consequence of that thesis). But it does not cut things quite finely enough for our purposes here. The thesis would be denied by a materialist who holds that it is positively conceivable that there be zombies (or other worlds physically identical to ours and phenomenally distinct), but that zombies are not metaphysically possible. (I have elsewhere called this view type-B materialism, as opposed to type-A materialism on which zombies are not even conceivable.) According to such a materialist, phenomenal truths about the world are not implied by the complete fundamental truth about the world, which is microphysical, so strong scrutability is false. This sort of denial of strong scrutability raises issues somewhat distinct from those I am concerned with here. For now, I am concerned with potential gaps between negative and positive conceivability, but this denial is best assimilated to a potential gap between positive conceivability and possibility. So it is useful to factor out a weaker scrutability thesis which this denial does not contravene.

The weaker scrutability thesis requires a sense of "complete qualitative description" such that on the type-B materialist view, a microphysical description is not a complete qualitative description. Intuitively, the microphysical description seems incomplete (in a sense) as a qualitative description precisely because it does not specify the phenomenal truths, and the phenomenal truths seem to be (in a sense) qualitative truths. And the sense in which these are qualitative truths seems to correspond to the fact that such truths will be required for a fully clear and distinct conception of what the world is qualitatively like. That is, they are required for a description of the world to the limits of positive conceivability.

A qualitatively complete description of the world, then, should be understood as a description to the limits of positive conceivability. That is, it is a description which specifies a unique positively conceivable situation. We can define this more precisely as follows:

A statement *D* is *qualitatively complete* if (i) *D* is positively conceivable, and (ii) if *D* & *F* is positively conceivable, then *D* implies *F*.

On the type-B materialist view, a complete microphysical description of the world will not be qualitatively complete. For various phenomenal truths *F*, *D* & *F* will be positively conceivable, as will *D* & $\sim F$, but *D* will imply neither *F* nor $\sim F$. On this view, a complete qualitative description of the world will require at least something akin to a full microphysical and phenomenal description.

With the notion of qualitative completeness in hand, we can now formulate our final version of the scrutability thesis.

SCRUTABILITY: If *D* is a qualitatively complete truth, then for all *S*, $S \rightarrow (D \text{ implies } S)$.

Or equivalently:

SCRUTABILITY If *D* is a qualitatively complete truth, then *D* is epistemically complete.

We can say that *S* is an *inscrutable truth* if *S* is true and some qualitatively complete truth *D* does not imply *S*. The scrutability thesis above says that there are no inscrutable truths.

It is easy to see that if *S* is inscrutable, then both *D* & *S* and *D* & $\sim S$ (for a relevant *D*) are in the twilight zone. *D* & *S* and *D* & $\sim S$ are negatively conceivable since their negations ($D \rightarrow \sim S$, $D \rightarrow S$) are not a priori. But they are not positively conceivable: if they were, then *D* could not be qualitatively complete. So if there are inscrutable truths, there are inhabitants of the twilight zone, and negative conceivability implies neither positive conceivability nor possibility.

Are there any likely candidates to be inscrutable truths? To assess this, it helps to have an idea of what a qualitatively complete truth will involve. It presumably will require at least microphysical information, including microphysical laws. It may or may not require phenomenal information (a type-A materialist view will deny this), but it cannot hurt to include it (specified in the sort of "pure phenomenal" vocabulary discussed by Chalmers 2002). On any of the type-A materialist, type-B materialist, and property dualist views, a qualitatively complete truth will imply the complete microphysical and phenomenal truth, so anything not implied by the former will not be implied by the latter. Indexical information is required, since more than one conjunction of complete objective truths with indexical claims will be positively conceivable. Finally, in order to rule out situations containing extra nonphysical, nonphenomenal goings-on, qualitative completeness requires a "totality" claim, holding that the world is a minimal world that satisfies the physical, phenomenal, and indexical claims specify.

Elsewhere (Chalmers and Jackson forthcoming), this conjunction of microphysical, phenomenal, indexical, and totality claims is referred to as PQTI. It is not entirely implausible that PQTI is itself a complete qualitative description of the world: for the most part, even where there are candidates M for truths not implied by PQTI, these do not seem to be associated with intuitions of distinct positively conceivable situations (verifying PQTI&M and PQTI&~M) analogous to the intuitions in the zombie case. In any case, even if PQTI is not itself qualitatively complete, we can at least say that any inscrutable truth will be a truth not implied by PQTI.

Possible candidates might fall into a number of classes.

(i) *Ordinary macroscopic truths.* One might first question whether ordinary macroscopic truths about the natural world, such as "grass is green" and "there is water in my pool" can be derived by a priori reasoning from PQTI. I have argued elsewhere (Chalmers 1996; Chalmers and Jackson forthcoming) that they can be, and I will not repeat that case here. But the basic idea is that straightforward a priori reasoning from PQTI puts one in a position to know all about the physical composition, the phenomenal appearance, the spatial structure, and the dynamic behavior of macroscopic systems, along with facts about their relation to oneself and their distribution in space and time; and this information in turns puts one in a position to know all ordinary macroscopic truths S about such systems, as long as one possesses the concepts involved in S. The information will include all the information on which ordinary perceptual or theoretical knowledge that S might be based, along with sufficient information to conclusively rule out skeptical counterpossibilities to S. If so, it is very plausible that PQTI implies S.

One worry arises with names and natural kind terms: someone might object that truths involving these terms cannot be a priori entailed by PQTI, as the relevant a priori connections are not built into the semantic content of these terms. In response, recall that we are working with a speaker-relative conception of apriority and primary intension. It may be in cases such as 'Neptune' or even 'water', the primary intension and a priori connections of a term varies between speakers, so that if "semantic content" must be common to all speakers, primary intensions and a priori connections are not determined by semantic content. But all that is required here is that the conditional "PQTI->S" be a priori for any given speaker. This thesis is quite compatible with the variation in primary intension, and it can be argued for straightforwardly along the lines of the previous paragraph.

(It might be thought that Kripke's epistemological arguments tell against even speaker-relative a priori entailments, but it is easy to see they have no power against the sort of specific entailment at issue here. At most, they suggest that a term's primary intension cannot be captured by a description. Special issues come up for expressions that are used with semantic deference, as when a speaker defers to other speakers in fixing a term's reference. I think that even these expressions can be accommodated on this framework, but for present purposes it is easiest to stipulate that we are concerned only with nondeferential uses.)

(ii) *Certain mathematical truths.* Someone might suggest that there are true mathematical statements that are not a priori, i.e. that are not knowable even on ideal rational reflection. For example, one might suppose that certain Gödelian statements in arithmetic (the Gödel sentence of the finite human brain?), or certain statements of higher set theory (the continuum hypothesis or its negation?) may be determinately

true without being ideally knowable. If such truths exist, they will plausibly not be implied by a qualitatively complete description of the world, so they will be inscrutable.

However, it is not at all clear that such statements exist. In any given case, one can argue that either (i) the statements in question are knowable under some idealization of rational reasoning, or (ii) that the statements are not determinately true or false. In the arithmetical case, one can argue that for any statement *S* there is some better reasoner than us that could know *S* a priori. Our inability to know a given Gödel sentence plausibly results from a contingent cognitive limitation: perhaps our limitations in the ordinal counting required for repeated Gödelization (which can be shown to settle all truths of arithmetic), or even our contingent inability to evaluate a predicate of all integers simultaneously (Russell's "mere medical impossibility"). In the case of unprovable statements of set theory, it is not at all clear that truth or falsity is determinate. Most set theorists seem to hold that the relevant cases are indeterminate (although see Lavine 1999 for an argument for determinacy); and even if they are determinate in some cases, it is not out of the question that possible beings could know the truth of further axioms that settle the determinate truths.

There is more to say about this issue. I think that the mathematical case is the most significant challenge to scrutability, and even if it fails, it clearly raises important questions about just what sorts of idealizations are allowed in our rational notions. For now, however, it suffices to note that there is no strong positive reason to hold that cases of mathematical determinacy without apriority exist.

(iii) *Vague statements*. It is plausible that some statements involving vague predicates (e.g., "person *X* is bald") cannot be known a priori to be true or false, even given complete qualitative information. Complete qualitative information will tell us how much hair a person has, but may leave the question of tallness unsettled. On the standard view of vagueness, this will not be a case of inscrutability, since the statements themselves will be neither true nor false. On the epistemic theory of vagueness, however, such statements are determinate even if we cannot know their truth-values: there is a precise border between the bald and the nonbald, but we cannot know where it is. Some versions of the epistemic theory may hold that this is due to rational limitations on our part (so that more intelligent creatures might be able to locate the border between baldness and nonbaldness); but we can consider a version on which even this epistemic connection fails. On such a view, some statements involving vague predicates will be inscrutable truths.

For example, let us assume for simplicity that baldness supervenes on number of hairs. On this version of the epistemic theory, some truths of the form "*X* is bald" will not be implied by truths of the form "*X* has *n* hairs". Here, "*X* has *n* hairs and is not bald" will be ideally negatively conceivable, but impossible. Here, the two statements "*X* has *n* hairs and is not bald" nor "*X* has *n* hairs and is bald" are negatively but not positively conceivable. When we consider any imagined situation in the vicinity - one in which *X* has *n* hairs of a specified nature and has further qualitative properties that are inessential here - it verifies neither "*X* is bald" nor "*X* is not bald". So these statements fall into the twilight zone.

(Contrast the zombie case in the philosophy of mind, where it seems that "physical structure *P* and conscious" and "physical structure *P* and not conscious" are each positively conceivable, verified by two different modally imaginable situations. In the baldness case, and other cases of vagueness, there are no

two such distinct modally imaginable situations: at best, there are two coherently entertainable descriptions. So in these cases, unlike the zombie case, there is no call to include information about baldness explicitly in a qualitatively complete description of the world; the existing description was already qualitatively complete, at least as far as the matters here are concerned. So the truths about baldness fall into the gap between negative and positive conceivability (yielding inscrutability), whereas truths about consciousness do not.)

Of course all this is contingent on the truth of the epistemic theory of vagueness, and the epistemic theory is widely regarded as very implausible. In fact one might trace the implausibility of the epistemic theory at least in part to the way it denies inscrutability. In these cases, it seems that a subject has all the qualitative information that could possibly be relevant, and it seems almost obvious that given that information, the subject is in a position to know all there is to know about baldness here. So it might be argued that the intuitive implausibility of the epistemic theory is grounded in an intuitive endorsement of scrutability, at least in this domain.

One might worry about the status of vague statements even when the epistemic theory is rejected. Here the answer depends on one's view of vagueness. If one accepts the law of the excluded middle ($S \vee \sim S$), then one must also accept $(PQTI \text{ implies } S) \vee (PQTI \text{ implies } \sim S)$. If one rejects the law of the excluded middle (the best option, on my view), then one will reject the corresponding claim about implication. If there are determinacy operators "det" and "indet", and cases in which $\text{indet}(S)$, then PQTI must imply $\text{indet}(S)$. But I do not think there are any fatal problems for scrutability here, over and above the problems that arise in analyzing vagueness in general.

(4) *Moral claims.*

Many philosophers hold that the truth or falsity of moral claims, such as "eating animals is bad", is not determined a priori by natural truths. This can be argued by a generalization of Moore's open question argument, suggesting that two people possessing full natural information might disagree on the truth of a claim like this, without either displaying incapacities of reasoning or failing to grasp moral concepts. This view is often combined with the view that moral claims are not strictly true or false at all; but some philosophers hold that moral claims are true or false despite being epistemically underdetermined by the natural truth in this way. On this view, moral truths are not implied by PQTI. As in the case of vagueness, there do not seem to be distinct positively conceivable situations verifying $PQTI \& M$ and $PQTI \& \sim M$, where M is a moral claim. If so, then moral truths are inscrutable truths.

As in the case of vagueness, this view of morality is controversial, so it certainly does not provide a clear case for inscrutable truths. Proponents of this sort of view often argue for the view by appealing to Kripke's distinction between the a priori and the necessary, but there are strong disanalogies: Kripke's cases are compatible with an entailment from (negative) conceivability to possibility, and with the scrutability of truth, whereas this view is not. And I think that there are good arguments against the view, based on considerations related to scrutability (Horgan and Tienson 19xx give some related arguments). But in any case, the view helps to illustrate what an inscrutable truth might be.

(A related issue: Yablo (this volume) worries that in the moral case, given the nonapriority of $PQTI \rightarrow M$ and $PQTI \rightarrow \sim M$ and a principle connecting nonapriority of $\sim S$ with primary possibility of S , one could infer the primary possibility of both $PQTI \& M$ and $PQTI \& \sim M$, which seems wrong. In response: if one denies the apriority of the conditionals and also rules out the position above, then the most plausible position remaining is that moral claims are not strictly true or false but are indeterminate; and it is plausible that if this view is true, it is a priori. If so, then it is a priori that $\sim \det(PQTI \& M)$ and that $\sim \det(PQTI \& \sim M)$. If so, then (given the official characterization of ideal negative conceivability) $PQTI \& M$ and $PQTI \& \sim M$ are not ideally negatively conceivable, so they will not be primarily possible.)

(5) *Metaphysical claims.*

Many issues within philosophy are such that it is not obvious that they can be conclusively settled by rationally reasoning from the information in PQTI. This applies especially to questions at issue within metaphysics: Do mereological sums exist? Do all dispositions have categorical bases? Are properties universals or tropes? What is required for identity over time? Do numbers exist? Is an A-theory or a B-theory of time correct? And even: does conceivability suffice for possibility?

There is no space here to consider these issues separately. But in general, an advocate of scrutability can take one of three strategies any given one of these issues. (1) Argue that sufficient rational reflection - perhaps more than has been done to date - can conclusively settle the issue (perhaps the issues about numbers and conceivability fall here). (2) Argue that the issue is positively conceivable either way, so that PQTI needs to be supplemented by further information to yield a qualitatively complete description of the world (perhaps the issues about categorical bases and about time fall here). (3) Argue that there is no fact of the matter about the issue, or that it can only be settled by terminological refinement (perhaps the issues about mereological sums and identity over time fall here). In each case, there is room for argument, but it certainly seems that there are no clear cases where all three of these strategies fail.

Overall: It seems that there is no clear counterexample to scrutability. At best there are some unclear potential counterexamples, none of which carries enormous antecedent plausibility, although some deserve further investigation.

Stepping back for a moment, why should we accept the scrutability thesis? One way to argue for it is to suggest that any reasonable candidate for an inscrutable truth will be an unknown and unknowable truth, since what we know is limited to information gained through perception (and so present in or implied by PQTI), plus that derived from rational reflection (and so implied by PQTI). One can also argue that all unknown truths stem from either ignorance of the qualitative nature of the world, or from insufficient a priori reasoning; and that the only unknowable truths stem from ignorance of the qualitative nature of the world. If this is so in general, then there are no inscrutable truths.

None of this yields a knockdown argument, but it does give reason to take the scrutability thesis very seriously.

Sideline: Modal Rationalism and Logical Empiricism

As Yablo (this volume) notes, the scrutability thesis has something in common with some versions of logical empiricism. A common logical empiricist thesis was that phenomenal truths analytically entail all truths. The main differences between the theses are (1) in appealing to a qualitatively complete truth, the scrutability thesis allows much more in its entailment base than just phenomenal truths (in particular, it allows the complete microphysical truth); and (2) scrutability as I have characterized appeals not to analyticity but to apriority, which I think is a more basic notion.

Yablo suggests that the problems of logical empiricism may infect the modal rationalism that I advocate. The problem on which Yablo focuses is the underdetermination of theory by evidence. Underdetermination of theory by *local* evidence is no problem for a sufficiently holistic logical empiricist, but underdetermination of theory by *total* evidence is a real problem. It seems easily conceivable that different states-of-affairs might provide the same evidence, or that some truths might leave no trace on our evidence; and there are even pairs of real-life scientific theories (as in quantum mechanics) that save all appearances while making different microphysical claims. These considerations are all tied to the limitations of observation, however: they suggest that phenomenal truths or observational truths underdetermine theory. They do nothing at all to suggest that the complete qualitative truth (including microphysical truths) underdetermines theory. So the most common worries about underdetermination do nothing to threaten the scrutability thesis.

To make a stronger parallel argument, Yablo appeals to a different worry, about the role played by a posteriori considerations of "reasonableness" and "sensitivity" in moving from evidence to theory. There seem to be a number of separate arguments here, although I am not certain that Yablo intended all of them.

First: considerations of reasonableness cannot be reduced to a set of a priori rules (as in "the dream of an a priori inductive logic"). Response: there is no reason to suppose that a priori truths, or a priori entailments, must be reducible to some basic set of explicit formal principles. So this is a red herring. (See Chalmers and Jackson forthcoming for more on this.) At best, this point might affect a thesis cast in terms of analyticity, if analyticity is defined in terms of logic plus definitions.

Second: the modal rationalist requires too much of our "grasp of meaning", by requiring knowledge of the relevant conditionals. Response: On the rationalist view, knowledge of the relevant conditionals need not be built into grasp of meaning, and need not be possessed by every subject who possesses the relevant concepts. Where present, the knowledge is usually a product of substantive reasoning, grounded both in possession of the relevant concepts and in rational reflection.

Third: there are cases where rational reflection on qualitative information underdetermines theoretical truth, which is settled only by pragmatic factors. Response: if the pragmatic factors are rationally underdetermined, then these are cases where originally indeterminate statements become determinately true or false because of terminological evolution. (See Chalmers and Jackson forthcoming.) So there is no point in the process at which there are inscrutable truths.

Fourth: the inference from PQTI to macroscopic truths may depend on "peeking", as when one perceptually imagines the appearance of a situation. Response: This point might threaten a scrutability thesis based on PTI, as information about appearances may be crucial but disallowed. But given the phenomenal information Q about appearances (in a pure phenomenal vocabulary), peeking comes for free. Knowledge of PQTI yields knowledge of the phenomenology of the appearances, and this puts one in as good a position to reason from those appearances to macroscopic truths as if one had experienced the appearances directly. (See also the further discussion below.)

Fifth: the general exercise of "sensitivity" is a posteriori, since it involves introspective knowledge of concept application in one's own mind. Response: No introspective knowledge is needed to know the entailment from PQTI to S. One merely needs to deploy the concepts involved in S; one does not need to observe their deployment.

There is more to say here, but it seems that at least on a first look, the scrutability thesis is unthreatened by its parallels with logical empiricism. Still, the parallels certainly exist, both with logical empiricism and other broadly phenomenalist and anti-realist views. These anti-realist views hold (in a sense) that when it comes to truth, nothing is hidden. The scrutability thesis does not suggest this, and is perfectly compatible with a realist view. But it suggests a weaker thesis: *given* complete qualitative knowledge (and ideal reflection), nothing is hidden. There is perhaps a tiny residue of anti-realism here: if a claim cannot be settled by a priori reasoning on the basic qualitative facts, then there is no fact of the matter about that claim. But one does not have to be a logical positivist to find this thesis attractive. (After all, there are some things that it is perfectly reasonable to be anti-realist about.) Indeed, in almost all cases where a claim cannot be settled in this way, as we have seen, it is independently plausible that the claim is indeterminate. So the scrutability thesis may well embody a principle that is tacit in our reasoning.

THE STRUCTURE OF THE TWILIGHT ZONE

We have seen that if there are inscrutable truths, then there are inhabitants of the "twilight zone" between negative and positive conceivability. We have also seen that there is no clear reason to accept that there are inscrutable truths. Where does this leave us on the relation between negative and positive conceivability more generally? Are there other inhabitants of the twilight zone?

To address this, it is useful to generalize the scrutability thesis slightly, so that it applies not only to complete qualitative descriptions of the actual world and to actual truths, but to any complete qualitative descriptions and to any truths. The generalized thesis is that a complete enough qualitative description of *any* world leaves no truth about that world epistemically open. After all, it would be odd if scrutability turned out to be true in this world but not in others; the thesis seems to have a much more general source than that.

GENERALIZED SCRUTABILITY: If D is qualitatively complete, then D is epistemically complete.

Clearly generalized scrutability implies scrutability: the earlier scrutability thesis is the special case where D is a qualitatively complete *truth*. Scrutability does not logically imply generalized scrutability, but it is natural to think that if scrutability is true, generalized scrutability is probably true. If scrutability holds, it seems unlikely that it holds accidentally, because of the character of the actual world. Rather, its truth would seem to reflect something deep about concepts, truth, and reason. If this is right, then the two theses are likely to stand and fall together.

OPEN INCONCEIVABILITIES:

The second potential source of a gap between negative and positive conceivability arises from states of affairs that are *inconceivable*, but that are nevertheless not ruled out a priori.

There are quite likely many *prima facie* inconceivabilities: a rich source is provided by statements about phenomenal properties quite distinct from our own. For example, the claim that there are creatures with 12-dimensional phenomenal color spaces cannot be ruled out a priori, but it may be beyond our capacity to conceive of a situation verifying this claim. Such a conception might require phenomenal concepts (and ultimately phenomenal experiences to ground those concepts) that we simply lack. If so, such a claim is *prima facie* negatively conceivable, but not *prima facie* positively conceivable. This is not obviously a case of ideal inconceivability, however. We have already seen that the inconceivability here stems from a lack in our repertoire of phenomenal concepts, and this limitation is contingent. If we idealize away from this conceptual lack, then the situation in question will plausibly turn out to be conceivable after all. Presumably there are possible creatures with the relevant concepts, and such creatures would have no difficulty in conceiving of the situations in question.

Still, perhaps there are some features of the world, or of some world, that simply cannot be positively conceived at all. One example might be provided by intrinsic properties that are not phenomenal properties, and are not conceptually related to them. One might argue that the only way to form a conception of an intrinsic property is by direct acquaintance, as in the phenomenal case, or perhaps by a priori reasoning from concepts of intrinsic properties one has direct acquaintance with; think of the missing shade of phenomenal blue. (Of course one might form an extrinsic conception of an intrinsic property, such as "the property that is causally responsible for such-and-such", but this is not good enough here, as such a conception leaves open multiple epistemic possibilities as to the nature of the property.) And one might argue that the only intrinsic properties any subject can be directly acquainted with are phenomenal properties. If so, then any intrinsic properties that are not phenomenal properties will be in the relevant sense inconceivable.

Of course all the assumptions going into the case above are highly contestable, but the possibility of inconceivable features of the world does not seem easy to rule out. This can be exploited to yield perhaps the most plausible example of an open inconceivability: namely, "there are inconceivable features of the world". This statement is by its nature verified by no positively conceivable situation, but it is also not easy to rule out a priori. Unless some way can be found to rule out this statement a priori, it will be (ideally) negatively conceivable but not positively conceivable, and hence will be in the twilight zone.

More precisely:

S is an **OPEN INCONCEIVABILITY** if S is negatively conceivable, but for all qualitatively complete D, D implies $\sim S$.

(Note: to handle indeterminacies, the last clause should hold that for all D, D implies $\sim \text{det}(S)$. If S is negatively conceivable, $\text{det}(S)$ is not ruled out a priori; if nevertheless for all D, D implies that S is indeterminate, then S should be an open inconceivability.)

We have seen that "There are no inconceivable features of the world" is one potential open inconceivability. Another is "There is no PC-complete description of the world". At a more specific level, the case of nonphenomenal intrinsic properties, on the assumptions above, will provide examples of open inconceivabilities insofar as there are ways to express relevant inconceivable truths (e.g., "There are nonphenomenal intrinsic properties", if nothing else). Still, none of these yield clear cases, so we can at least formulate a relevant thesis opposed to them:

NOINCONCEIVABILITY: No S is an open inconceivability.

It is not clear how best to argue for this thesis. One might argue for any property, there is some creature than can form a conception of it - perhaps any intrinsic property can be known by acquaintance, and any non-intrinsic property by description. And one might argue that this principle is itself a priori. If that is so, then it plausibly follows that there are no open inconceivabilities. But the central claim here is far from obvious.

Like inscrutabilities, open inconceivabilities (if they exist) provide a gap between negative and positive conceivability. They differ, however, in that they may not provide a gap between negative conceivability and possibility. That gap depends on whether the open inconceivabilities in question correspond to real possibilities (e.g. properties we can't form a conception of), or whether they correspond to impossibilities that we cannot rule out a priori (perhaps all properties are conceivable, but we can't rule out the alternative a priori). If all open inconceivabilities are of the former sort, then negative conceivability might still be a guide to possibility - it is just that the possible will outstrip the positively conceivable. If some are of the latter sort, then negative conceivability will be an imperfect guide to possibility.

TWILIGHT ZONE SUMMARY:

We have seen that potential members of the twilight zone stem from at least two classes: inscrutabilities and open inconceivabilities. In fact it is not hard to see that all members of the twilight zone stem from these two classes.

NEGPOS: Ideal negative conceivability entails ideal primary positive conceivability.

The negpos principle says that there are no members of the twilight zone. Then

Claim: NEGPOS $\leftrightarrow$ GENERALIZED SCRUTABILITY & NOINCONCEIVABILITY

Proof: Left-to-right: Given negpos, any qualitatively complete statement will be epistemically complete, so generalized scrutability will be true. Given negpos, any negatively conceivable statement will be positively conceivable, so will be entailed by some qualitatively complete D, so noinconceivability will be true. Note that the second part requires the principle that any positively conceivable statement is implied by some qualitatively complete statement. This seems reasonable, as it encapsulates the idea that a statement verified only by "uncompletable" situations will not be ideally positively conceivable.

Right-to-left: Let S be negatively conceivable. Noinconceivability implies that there is a D such that $\sim(D \text{ implies } \sim \text{det}(S))$. Generalized scrutability implies that D settles S's truth-value, so D implies S. (Note that the determinacy operator in the definition of open inconceivability is needed here, in order to exclude the possibility that D implies neither S nor $\sim S$, due to indeterminacy.)

So in order to close a potential gap between (ideal primary) negative and positive conceivability, it is necessary and sufficient to rule out generalized inscrutabilities and open inconceivabilities.

In order to close a potential gap between negative conceivability and possibility, it is necessary to rule out generalized inscrutabilities (if S is a generalized inscrutability, then both $D \rightarrow S$ and $D \rightarrow \sim S$ will be negatively conceivable for a relevant D, but both cannot be possible). It is not necessary to rule out all open inconceivabilities, but one must rule out all *impossible* open inconceivabilities. If some open inconceivabilities are also impossibilities, then negative conceivability does not entail possibility. But if all open inconceivabilities are possibilities (which is not entirely implausible), then an entailment between negative conceivability and possibility is not threatened.

FROM POSITIVE CONCEIVABILITY TO POSSIBILITY

Does (ideal primary) conceivability imply (primary) possibility? A counterexample to this principle must involve what I have elsewhere called a *strong necessity*: a statement that is falsified by some positively conceivable situation (considered as actual), but which nevertheless true in all possible worlds (considered as actual). For such necessities to exist, the space of positively conceivable situations must outstrip the space of possible worlds.

There are certainly no clear examples of strong necessities, and the only candidates are highly tendentious. I have discussed this matter at some length elsewhere (Chalmers 1999), so I will say only a little about some possible candidates here.

(i) *The existence of God*. On many theist views, a god exists necessarily, so that every possible world contains a god. But a theist may hold that it is not a priori that a god exists, and that a godless world is positively conceivable, even on rational reflection. On this view, it is natural to hold that "no god exists" is primarily positively conceivable, but not primarily possible. So if this view is correct, then "a god exists" is a strong necessity. Of course, the theist thesis here is highly controversial, so this is not a strong

counterexample, but it illustrates what a counterexample must involve.

One further worry: if god does not exist necessarily, then "A necessary god exists" is impossible. But it may seem that a necessary god is at least conceivable (see Yablo 1999). In response, I deny that a necessarily existing god is ideally conceivable. A god's existence may be conceivable, but to conceive of a god's necessary existence is much harder, especially given its conceivable nonexistence. In effect, one must conceive (metamodally!) that conceivability does not imply possibility. But it is not clear that this is more than *prima facie* negatively conceivable. On my view, it is *a priori* (if non-obvious) that conceivability entails possibility (see below for the sketch of an *a priori* argument). If so, then the denial of the entailment is not ideally conceivable, and so neither is the necessary existence of a god.

(2) *Laws of nature*. Some philosophers hold that the laws of nature are metaphysically necessary. On some views of this sort (e.g. those discussed by Fine and Sidelle in this volume), this necessity arises for broadly Kripkean reasons: the reference of terms such as "mass" is fixed *a posteriori* to a certain very specific property, so that worlds with different laws do not contain mass. I think this view is implausible, but in any case it is compatible with an entailment from primary conceivability to primary possibility. If *G'* is a counternomic statement (say, an adjusted statement of gravitational laws with a different constant), then *G'* is both primarily conceivable and primarily possible. *G'* is verified by a metaphysically possible world *W* considered as actual, although not by *W* considered as counterfactual. (Considered as counterfactual, *W* contains "schmass", not mass.) So there are no strong necessities here.

There is a stronger view on which the laws of every world are the laws of our world, governing properties instantiated in our world. On this sort of view, even "schmass" worlds are metaphysically impossible: *G'* will be primarily conceivable but not even primarily possible. So on this view, laws of nature are strong necessities. There is no reason to accept this view, however. (It is notable that Fine and Sidelle quickly dismiss such a view as too extreme to be plausible.) Proponents of necessary laws usually appeal to Kripke's necessary *a posteriori* for support, but the Kripkean cases support at best the weak view in the previous paragraph. Nothing here gives reason to suppose that worlds with different laws are impossible; at best, it suggests that they are misdescribed as breaking our laws. So there is no good reason here to deny the conceivability-possibility thesis.

(3) *Response-enabled concepts*. Yablo (this volume) considers a class of *response-enabled* concepts whose extension, and whose meaning, is fixed by our responses. He suggests that "oval" is in this class: the reference of "oval" is fixed by picking out whatever looks oval-shaped to us, irrespective of any pure geometric description of their shape. I think the example is unfortunate, since "oval" is arguably a pure geometric concept, picking out certain geometric shapes regardless of the responses they cause in us. But there may be other terms that function as Yablo suggests, so I will play along with his suggestion that "oval" works this way.

Following Yablo, let "cassini" be a term for a certain class of mathematically defined geometric figures, of a sort that actually cause "oval"-responses. Then (given Yablo's view of "oval"), it is not *a priori* that cassinis are ovals. So "cassinis are not ovals" is ideally negatively conceivable. "Cassinis are not ovals" is also plausibly ideally positively conceivable, since it is verified by a situation in which cassinis are not the

sort of object that cause "oval"-responses. But Yablo suggests that "cassinis are ovals" is nevertheless true in all worlds considered as actual: in all such worlds, cassinis fall under the extension of "oval". If this is correct, then ideal (negative or positive) primary conceivability does not entail primary possibility.

In response: As I have characterized considering-as-actual, it is clear that "cassinis are oval" is true of some worlds considered as actual. Let W be a world in which cassinis cause oval-responses. Let us grant that it is a priori that if W is actual, cassinis cause oval-responses. (Yablo raises no objection to this.) We can also note that if "oval" functions as described, then a material conditional such as "if H s cause oval responses, then H s are ovals" is pretty clearly a priori. (Yablo himself allows that there may be an apriori connection here.) It follows that the material conditional "if W is actual, cassinis are ovals" is a priori. So as I have defined things, "cassinis are ovals" is true in W considered as actual, and is primarily possible.

Why does Yablo resist this straightforward conclusion? It seems to me that he is operating with a different conception of how statements are evaluated in considering a world as actual, one tied to the "if it had turned out" locution and to certain claims about "conceptual necessity". I am not sure that I fully grasp this conception, but for present purposes I need not deny that it is coherent or that it captures some feature of our concepts. But it is clearly distinct from the conception I am operating with, on which considering as actual involves a priori reasoning about epistemic possibilities. Yablo gives no reason to deny that this conception is coherent, or that it yields the results I have suggested. So the conceivability-possibility link that I have advocated is unthreatened by Yablo's discussion.

Something similar applies to other claims that Yablo discusses, including "unless we are greatly misled about the circumstances of visual perception, what looks green is green". If this (or something like it) is a priori, then like all a priori statements, it will automatically hold in all worlds considered as actual, at least on my conception of considering as actual (though perhaps not on Yablo's). And the statement " F s are not red", where " F " involves a complete intrinsic characterization of something that is actually red, will be a posteriori, primarily positively conceivable, and primarily possible on my conception: it will be straightforwardly true in a world (considered as actual) where F s do not look red.

At one point Yablo raises another worry about response-dependent concepts: physical and phenomenal truths may not imply truths about yellowness (say), since an characterization of the relevant phenomenology may not enable one to identify the "yellow" responses a priori. I think this is not a problem. The relevant responses are phenomenal kinds, characterized by what it is like to have them. In particular, knowledge of what it is like to experience an object (in normal circumstances) enables knowledge of whether the object is yellow, with no further empirical justification required. Further, physical and phenomenal knowledge enable knowledge of what it is like to experience the relevant objects (in normal circumstances), with no further empirical justification required. It follows that the physical and phenomenal truths imply the truths about yellowness.

(4) *Psychophysical laws*. A final example is given by some type-B materialist views, on which there is an epistemic gap between the physical and the phenomenal, but no ontological gap. On such a view, zombies (and the like) are positively conceivable but not possible. Type-B materialists often appeal to Kripkean cases for support, but it is not hard to see that these do not help, since those cases are compatible with the

primary conceivability-possibility link, and even the primary possibility of zombies causes problems for materialism. In response, some type-B materialists deny that zombies are even primarily possible. On such a view, psychophysical laws (of the form "if P, then Q" for physical P and phenomenal Q) are strong necessities.

Again, this view is highly controversial, so it does not provide any clear counterexample. This view is usually put forward on the grounds that it is the only tenable way to preserve materialism, given the epistemic gap; but of course that falls well short of a positive argument for the view, especially when the truth of materialism is at issue. Indeed, one can suggest that the conceivability-possibility link that holds elsewhere itself provides a strong argument against this view. Given the discussion above, it seems that the strong necessities required here will be unique. (Even if one thinks that God and laws of nature provide partners in crime, it is notable that the sort of strong necessity at issue there cannot save materialism: in those cases, strong necessities connect ontologically distinct existences!) Some type-B materialists (e.g. Loar 1997 and Hill 1998) have bitten this bullet and tried to give an *explanation* of why strong necessities should uniquely arise in the phenomenal domain. I have argued elsewhere (Chalmers 1999) that these explanations fail.

In summary: in each case, the claim that there are strong necessities rests on very controversial assumptions. One might more plausibly argue in reverse: in each of these cases, the elsewhere unbroken link between conceivability and possibility provides an argument against the assumptions in question. In any case, there is no strong threat to the conceivability-possibility thesis here.

Still, all this at best makes a negative case for the conceivability-possibility thesis, by defeating potential counterexamples and explanations. It remains to make a positive case for the thesis, giving reasons why we should expect it to be true. I make a start on this case in Chalmers (1999). I hope to expand on this further elsewhere, but here I will just recapitulate the case briefly.

The argument involves locating the roots of our modal concepts in the rational domain. When one looks at the purposes to which modality is put (e.g. in the first chapter of Lewis 1986), it is striking that many of these purposes are tied closely to the rational and the psychological: analyzing the contents of thoughts and the semantics of language, giving an account of counterfactual thought, analyzing rational inference. It can be argued that for a concept of possibility and necessity to be truly useful in analyzing these domains, it must be a *rational* modal concept, tied constitutively to consistency, rational inference, or conceivability.

It is not difficult to argue that even if not all conceivable worlds are metaphysically possible worlds, we *need* a rational modal concept -- call it logical possibility - to best analyze the phenomena in question. For example, even if all counternomic worlds are metaphysically impossible, it will still be tremendously useful to have a wider space of worlds (or world-like entities) to help analyze and explain the hypotheses and inferences of a scientist investigating the laws of nature. Further, there is no bar to a space of such worlds. If one does not want simply to postulate them, one can easily construct them in an "ersatz" way - via equivalence classes of qualitatively complete (or epistemically complete) descriptions, for example - and introduce means of semantic evaluations such that the worlds in question are perfectly well-behaved.

One can then argue that this space of worlds suffices to account for all modal phenomena that we have reason to believe in. Such a space will analyze such rational and psychological matters as counterfactual thought, rational inference, and the contents of thought and language as well as any other modal can. And with the help of a little two-dimensional semantics, it can account for such "metaphysical" modal phenomena as the concept/property distinction, a posteriori necessities, and so on. The two-dimensional semantics in question will be grounded in a priori conceptual analysis plus nonmodal facts about the actual world. So one modal space plus conceptual analysis plus nonmodal facts gives us everything, as long as this modal space is tied constitutively to the rational domain. If this modal space is all, we have *modal monism*.

The believer in strong necessities, by contrast, must embrace a *modal dualism*, with distinct primitive modalities of logical and metaphysical possibility, neither of which is reducible to the other. There is no good reason to accept such a modal dualism, when modal monism can explain all the untendentious phenomena. One can argue further that there is no distinct *concept* of metaphysical possibility for the second modality to answer to. The momentary impression of such a concept stems from a confused understanding of such ontic/epistemic distinctions such as that between apriority and necessity, and that between concept and property, all of which are easily subsumed under a modal monism with the help of some two-dimensional semantics.

Ultimately, there is just one circle of modal concepts, including both the rational modal concepts (validity, rational entailment, a priority, conceivability) and the metaphysical modal concepts (possibility, necessity, property). The result we are left with is *modal rationalism* in more senses than one: a priori access to modality, and constitutive ties between the modal and rational domains.

CONCLUSIONS

We can sum up the lay of the land by labeling some varieties of modal rationalism:

WEAK MODAL RATIONALISM: (Ideal primary) positive conceivability entails (primary) possibility.

STRONG MODAL RATIONALISM: Negative conceivability entails possibility.

PURE MODAL RATIONALISM: Positive conceivability $\leftrightarrow$ negative conceivability $\leftrightarrow$ possibility.

Then

PURE MODAL RATIONALISM $\leftrightarrow$ WEAK MODAL RATIONALISM & NEGPOS

The left-to-right direction here is obvious, and the right-to-left direction follows from the observation that

possibility entails negative conceivability (no primary possibility is ruled out a priori). Combining this with the previous result about the nature of the twilight zone, we can say:

PURE MODAL RATIONALISM <-> WEAK MODAL RATIONALISM &
GENERALIZED SCRUTABILITY & NOINCONCEIVABILITY.

It follows that to establish pure modal rationalism, we must rule out strong necessities, generalized inscrutabilities, and open inconceivabilities. Here I am most confident about the first, reasonably confident about the second, and unsure about the third. I have outlined a case against strong necessities here, and given tentative reasons to be doubtful about inscrutabilities, while the status of open inconceivabilities is unclear. In any case, it seems to me that each of these three is a distinct and substantial philosophical project, and that the investigation of each raises deep philosophical questions and promises significant philosophical rewards.

If weak modal rationalism is the best we can establish, then we will have done enough to support conceivability arguments as traditionally used, although the overall picture of modality and of modal epistemology will remain somewhat messy. We will have distinct notions of positive and negative conceivability, and thus a mild dualism within the rational modal sphere (though of course it will be a dualism that is forced on us by the phenomena). If there are generalized inscrutabilities, then although conceivability will guarantee access to a possible world, it may not yield access to all truths in that world. And if there are open inconceivabilities, there will be worlds that conceivability offers no access to.

Pure modal rationalism yields a simpler picture of modal space, and a correspondingly elegant epistemology. Looking at its three components in turn: the first says that positive conceivability gives us access to only possible worlds, the third says that it gives us access to all the possible worlds, and the second says that we can know all the truths about these possible worlds. In effect, we have a telescope that gives us access to all and only the stars, and that tells us the exact composition of every star. If this thesis is true, the epistemology of modality, at least when idealized, will be simple and beautiful.

APPENDIX: THE MIND-BODY PROBLEM

With these conceivability-possibility theses in hand, it is interesting to apply them to various conceivability arguments against materialism in the philosophy of mind.

Historically, the most important such argument has been Descartes' conceivability argument. This argues from the conceivability of my existing without a body to the possibility of my existence without a body, and so to the claim that I am not physical. The soundness of this argument is often doubted, and the standard reasons for doubt can be expressed straightforwardly in the current framework. The sense in which it is clearly conceivable that I am disembodied is primary positive conceivability, from which the 1-possibility of disembodiment follows. The sense in which physical things are essentially physical involves 2-necessity (as do all claims of de re necessity). But the 1-possibility of disembodiment is quite compatible with the 2-impossibility of disembodiment, so the claim that I am physical is not threatened by Descartes' argument.

More recently, the knowledge argument and the zombie argument against materialism have been widely discussed. Here, let P be the conjunction of physical truths about the world, and let Q be a phenomenal truth. The zombie argument claims that zombies, and therefore $P \& \sim Q$, are primarily positively conceivable. (Here, Q might be "someone is conscious".) The knowledge argument claims that Q cannot be derived a priori from P , so that $P \& \sim Q$ is primarily negatively conceivable. (Here, Q might be "someone is having a such-and-such experience".) From here, both step to the denial of materialism. If we use the current framework to analyze these arguments, a first pass might yield something like the following:

- (1) $P \& \sim Q$ is ideally primarily positively (negatively) conceivable.
- (2) If $P \& \sim Q$ is ideally primarily positively (negatively) conceivable. then $P \& \sim Q$ is primarily possible.
- (3) If $P \& \sim Q$ is primarily possible, materialism is false.
-
- (4) Materialism is false.

Here, premise (2) is a special case of one of the two main conceivability-possibility theses already outlined. It is notable that the zombie argument requires a weaker epistemic-modal premise here: the thesis connecting positive conceivability to possibility is weaker than the thesis connecting negative conceivability to possibility. (It requires excluding only strong necessities, not inscrutabilities.) This is offset to some extent by the fact that the knowledge argument requires a weaker epistemic premise: the claim that $P \& \sim Q$ is negatively conceivable is weaker than the claim that it is positively conceivable.

What of the epistemic premise (1)? This is widely although not universally accepted in the knowledge argument case, and to a somewhat lesser extent in the zombie case. A materialist might deny it in two ways: either by denying even the *prima facie* conceivability of $P \& \sim Q$, or by accepting *prima facie* conceivability but denying ideal conceivability. Some type-A materialists will deny even *prima facie* conceivability, but this denial is not easy to defend, since it runs counter to a very strong intuition. Others accept *prima facie* conceivability but deny ideal conceivability, holding that there may be a deep epistemic connection between P and Q , and a deep a priori contradiction in the notion of a zombie.

The second position, exploiting the gap between *prima facie* and ideal conceivability, may seem particularly promising, especially in the "type-C materialist" view that holds that there is a deep connection that we have not found yet (or perhaps cannot find). But there are two problems. First, this materialist will concede that zombies are not just *prima facie* but *secunda facie* positively conceivable, and we have seen that *secunda facie* conceivability is an extremely good guide to possibility. Second, defeating ideal conceivability will require an a priori entailment from physical to phenomenal, which will require an analysis of phenomenal concepts that can support that entailment. Given the structural-dispositional nature of the physical concepts in P , this could only be a structural or functional analysis. But there is good reason to believe that any such analysis of phenomenal concepts is a misanalysis. So while type-C strategy is an interesting strategy that deserves investigation, I think we have reason to believe that it will not succeed.

That leaves premise (3). Here, one runs up against the same problem as in the Cartesian argument. Materialism requires that the physical truths *secondarily* necessitate all truths, and so requires that $P \& \sim Q$ is secondarily possible. But there is no clear inference from the primary possibility of $P \& \sim Q$ to its secondary possibility. So the argument seems to be unsound as it stands. (An eliminativist who denies Q might deny (3) for different reasons, but I will set that position aside here.)

In this case, unlike the Cartesian case, the argument can be rescued. First, one can observe that *if* P and Q both had identical primary and secondary intensions (up to centering), then premise (3) would be straightforwardly true. Further, it is very plausible that the most important phenomenal concepts do indeed have the same primary and secondary intensions (see Chalmers forthcoming), so that Q at least can be accommodated here. And even if this is false, Q's primary intension can be seen as the secondary intension of some other truth Q', which stands to Q roughly as "watery stuff" stands to "water". As long as P has the same primary and secondary intension, then the primary possibility of $P \& \sim Q$ will entail the secondary possibility of $P \& \sim Q'$, which will itself entail the falsity of materialism.

A loophole emerges: it is not clear that P has the same primary and secondary intension. It can reasonably be argued that physical concepts have their reference fixed by some dispositional role, but refer to an underlying categorical property. If so, their primary intensions pick out whatever plays a certain role in the world (irrespective of categorical nature), while their secondary intensions pick out instances of a certain categorical property (irrespective of its role). If so, the purported "zombie world" in which the primary intension of $P \& \sim Q$ holds may be a world in which the secondary intension of P is false, so we cannot infer the secondary possibility of $P \& \sim Q$ (or $P \& \sim Q'$).

However, this loophole opens up only a small space for the materialist. Consider the conceived world W, in which the primary intension of $P \& \sim Q$ holds. Because the primary intension of P holds, this world must be structurally-dispositionally isomorphic to the actual world, with the same patterns of microphysical causal roles being played. If P's secondary intension fails, it can only be because these microphysical causal roles have different categorical bases in W (or just possibly, no categorical bases at all). This difference is the only microphysical difference between our world and W. If physicalism is true, it is this difference that is responsible for the presence of consciousness in our world and its absence in W.

What results is a view on which the existence of consciousness is not necessitated by the structural or dispositional aspects of the microphysics of our world, but is necessitated by the categorical aspects of microphysics (the underlying categorical basis of microphysical dispositions), perhaps in combination with structural/dispositional aspects. This is an important view: it is the view put forward by Russell (1926) and discussed in recent years by Maxwell (1978), Lockwood (1989), and others. In effect, the view holds that consciousness stems from the underlying categorical aspect of microphysics. On this view, the nature of the categorical aspect is left open by physical theory, but it turns out to involve special properties that are collectively responsible for constituting consciousness. We can call these special properties *protophenomenal*: they might not themselves be phenomenal properties, but they stand in a constitutive relation to phenomenal properties. We can call the view as a whole *panprotopsychoism*.

It is not clear whether this sort of panprotopsyism qualifies as a version of physicalism. That question turns on whether the underlying protophenomenal properties are best counted as physical properties, or not. We need not settle that question here: We need only note that if it is a sort of physicalism, it is a quite unusual sort, and one that many physicalists do not accept. In many ways, it has more in common with nonmaterialist views, in virtue of its postulation of fundamental protophenomenal properties whose nature is not revealed to us by physical theory.

In any case, we are now in a position to reformulate the relevant argument:

- (1) $P \& \sim Q$ is ideally primarily positively (negatively) conceivable.
- (2) If $P \& \sim Q$ is ideally primarily positively (negatively) conceivable, then $P \& \sim Q$ is primarily possible.
- (3) If $P \& \sim Q$ is primarily possible but not secondarily possible, then panprotopsyism is true.
- (4) If $P \& \sim Q$ is secondarily possible, materialism is false.
-
- (5) Materialism is false or panprotopsyism is true.

The argument (in both versions) is valid, and I have given reasons to accept all of the premises. Note that one can substitute the secondary possibility of $P \& Q'$ for the secondary possibility of $P \& Q$ in the third and fourth premises, if necessary. Note also that I have said nothing about the role of indexicals and centering. One might think that these raise another loophole in the argument (around premise (3)), by opening another gap between primary and secondary possibility. It is not hard to give a fuller version that takes this role into account (see Chalmers 1998), but I omit the details here for reasons of space.

Finally, a note on Stalnaker's paper in this volume, concerning the zombie argument. Stalnaker (through his character "Anne") questions the argument, by questioning the inference from conceivability to possibility. He invokes a notion of conceivability distinct from those discussed here, which we might call *1-2-conceivability*. It is 1-2-conceivable that S if it is primarily conceivable that S is secondarily possible, or more precisely, if "possibly S" is 1-conceivable, where the modal operator here represents 2-possibility. Stalnaker accepts that zombies are 1-2-conceivable (property dualism is not a priori false, and if property dualism is true, then zombies are 2-possible). But he notes that the 1-2-conceivability of S does not entail the possibility of S: the epistemic possibility of property dualism is compatible with the truth of materialism, and with the 2-impossibility (and 1-impossibility) of zombies.

I agree that the conceivability of zombies, in this sense, does not directly entail the falsity of materialism. But this sort of conceivability plays no role in the arguments I have given. What is relevant is simply the *1-conceivability* of $P \& \sim Q$. In the knowledge argument, the argument aims to directly establish the non-apriority of $P \rightarrow \sim Q$, and so the prima negative conceivability of $P \& \sim Q$. In the zombie argument itself, the claim is that it is conceivable that *in the actual world*, P holds but no-one is conscious. (Of course I know that I am conscious, but this is a posteriori knowledge; that issue can also be bypassed by considering only the epistemic possibility that P holds while *others* in the actual world are zombies.) That is, the claim is that $P \& \sim Q$ is primarily positively conceivable. Stalnaker says nothing to cast doubt on this claim (or the

analogous claim about negative conceivability), and he says nothing to cast doubt on the inference from primary conceivability to primary possibility. So his discussion leaves this argument untouched.

ACKNOWLEDGMENTS: I have presented this material at Arizona State University, Princeton University, and the Mighty Midwestern Metaphysical Mayhem conference at Notre Dame in summer 1999. Thanks to many on those occasions and elsewhere for discussion of these issues. Special thanks to Tamar Gendler, John Hawthorne, Bernie Kobes, and David Sosa for comments on the paper, and to Steve Yablo for discussion and correspondence.

REFERENCES

Benacerraf, P. 1965. What numbers could not be. *Philosophical Review* 74:47-73.

Benardete, J.A. 1964. *Infinity: An Essay in Metaphysics*. Oxford University Press.

Block, N. & Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review* 108:1-46.

Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D.J. 1998. Mind and modality. Lectures given at Princeton University.
<http://www.u.arizona.edu/~chalmers/papers/mm.html>.

Chalmers, D.J. 1999. Materialism and the metaphysics of modality. *Philosophy and Phenomenological Research* 59:473-96.

Chalmers, D.J. 2002. The content and epistemology of phenomenal belief. In (A. Jokic & Q. Smith, eds) *Aspects of Consciousness*. Oxford University Press.
<http://www.u.arizona.edu/~chalmers/papers/belief.html>

Chalmers, D.J. (forthcoming a). On sense and intension.
<http://www.u.arizona.edu/~chalmers/papers/intension.html>

Chalmers, D.J. (forthcoming b). The nature of epistemic space.
<http://www.u.arizona.edu/~chalmers/papers/espace.html>

Chalmers, D.J. & Jackson, F. (forthcoming). Conceptual analysis and reductive explanation.
<http://www.u.arizona.edu/~chalmers/papers/analysis.html>.

Davies, M. & Humberstone, I.L. 1981. Two notions of necessity. *Philosophical Studies* 58:1-30.

Evans, G. 1977. Reference and contingency. *The Monist* 62:161-89.

Fine, K. (this volume).

Hawthorne, J. 2000. Before-effect and Zeno causality. *Nous* 34:622-33.

Hill, C.S. 1997. Imaginability, conceivability, possibility, and the mind-body problem. *Philosophical Studies* 87:61-85.

Horgan, T. 1986. Psychologism, Semantics, and Ontology. *Nous* 20: 21-31.

Jackson F. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford University Press.

Lavine, S. 1999. *Skolem was Wrong*. Book manuscript.

Lewis, D. 1986. *On the Plurality of Worlds*. Blackwell.

Loar, B. 1997. Phenomenal states (second version). In (N. Block, O. Flanagan, and G. Güzeldere, eds) *The Nature of Consciousness: Philosophical Debates*. MIT Press.

Loar, B. 1999. On David Chalmers' *The Conscious Mind*. *Philosophy and Phenomenological Research* 59:465-72.

Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford: Blackwell.

Maxwell, G. 1978. Rigid designators and mind-brain identity. In (C.W. Savage, ed.) *Perception and Cognition: Issues in the Foundations of Psychology* (Minnesota Studies in the Philosophy of Science, Vol. 9). Minneapolis: University of Minnesota Press.

Menzies, P. 1998. Possibility and conceivability: A response-dependent account of their connections. In (R. Casati & C. Tappolet, eds) *Response-Dependence* (European Review of Philosophy, volume 3). CSLI Press.

Quine, W. 1961. *Word and Object*. MIT Press.

Russell, B. 1926. *The Analysis of Matter*. Kegan Paul.

Sidelle, A. (this volume). On the metaphysical contingency of laws of nature.

Stalnaker, R. (this volume). What is it like to be a zombie?

van Cleve, J. 1983. Conceivability and the Cartesian argument for dualism. *Pacific Philosophical*

Quarterly 64:35-45.

Yablo, S. 1993. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53:1-42.

Yablo, S. 1999. Concepts and consciousness. *Philosophy and Phenomenological Research*.

Yablo, S. (this volume). Coulda, woulda, shoulda.

Conceptual Analysis and Reductive Explanation

David J. Chalmers & Frank Jackson

[[Forthcoming in the *Philosophical Review*. This is in part a reply to Block and Stalnaker's "[Conceptual Analysis, Dualism, and the Explanatory Gap](#)".]]

1 Introduction

Is conceptual analysis required for reductive explanation? If there is no a priori entailment from microphysical truths to phenomenal truths, does reductive explanation of the phenomenal fail? We say yes (Chalmers 1996; Jackson 1994, 1998). Ned Block and Robert Stalnaker say no (Block and Stalnaker 1999).

A number of issues can be distinguished:

- (1) Is there an a priori entailment from microphysical truths to ordinary macroscopic truths?
- (2) If there is no a priori entailment from microphysical truths to phenomenal truths, does reductive explanation of the phenomenal fail?
- (3) If there is no a priori entailment from microphysical truths to phenomenal truths, is physicalism about the phenomenal false?
- (4) Is there an a priori entailment from microphysical truths to phenomenal truths?

We hold that the first three questions should be answered positively (with some qualifications to be outlined). Block and Stalnaker hold that the first three questions should be answered negatively. Their central strategy is to argue for a negative answer to the first question, and to use this conclusion to argue for a negative answer to the second and third. They argue that truths about water and about life, for example, are not entailed a priori by microphysical truths, but that this is no bar to the reductive explanation and physical constitution of water and of life.

In this paper, we will address Block and Stalnaker's arguments for a negative answer to the first three questions, while remaining neutral on the fourth. We will proceed by first giving an independent defence of a positive answer to the first question. This makes the ensuing reply to Block and Stalnaker more straightforward, and also makes the discussion accessible to those unfamiliar with the literature.

2 Clarifying the thesis

The initial thesis at issue is whether there is an a priori entailment from microphysical truths to ordinary macrophysical truths.[*] Before proceeding, this thesis needs to be clarified and qualified to meet some obvious objections.

*[[Horgan (1984) puts forward a version of this thesis under the name "cosmic hermeneutics", although he qualifies the thesis to allow a role for a posteriori identities involving rigid designators in inferring macroscopic truths from microscopic truths. Chalmers (1996), Jackson (1994; 1998), and Lewis (1994) argue for versions of the thesis.]]

On our usage, P entails Q when the material conditional $P \rightarrow Q$ is true: that is, when it is not the case that P is true and Q is false. An a priori entailment is just an a priori material conditional. For ease of usage, we will speak of a priori entailment as *implication*. On this usage, P implies Q when the material conditional $P \rightarrow Q$ is a priori; that is, when it is possible to know that P entails Q with justification independent of experience. On this usage, entailment is a nonmodal notion, while implication involves an epistemic modality. We will assume that there are some a priori truths; general skepticism about the a priori is beyond the scope of this paper.

Let P be the conjunction of microphysical truths about the world. Microphysical truths are truths about the fundamental entities and properties of physics, in the language of a completed physics. We can also stipulate that P includes the conjunction of the fundamental laws of physics. We will not engage the issue of what counts as "physics", but will stipulate that if there are any fundamental mental entities or properties, they are not part of physics. (This begs no important questions: if the mental is fundamental, it cannot be reductively explained.) We also will not engage issues arising specifically from the foundations of quantum mechanics, and will take it for granted that microphysical truths include or imply truths about the spatiotemporal position, velocity, and mass of microscopic entities.[*] So P will likely include or imply truths about the distribution of fundamental entities (perhaps particles, waves, and/or fields) in space-time, truths about their fundamental properties, and truths about the fundamental laws that govern them.

*[[To simplify, we can assume something like a Bohmian interpretation of quantum mechanics. The central claim can also be defended under other interpretations, but further complexities are involved.]]

Let M be a typical macroscopic truth concerning natural phenomena, such as water or life, outside the psychological, social, and evaluative domains. Some such truths include: 'Water boils at 100 degrees Centigrade'; 'water is H₂O'; 'water covers much of this planet'; 'life propagates through the replication of DNA'; 'there are many living beings'; and so on. The initial thesis at issue is that for this sort of M , P implies M . This thesis needs to be immediately qualified to handle some obvious loopholes.

(1) *Negative truths*. As described, P is compatible with the claim that there are further nonphysical entities or properties in the world: angels, perhaps, or epiphenomenal ectoplasm. So P does not imply such negative truths (if they are truths) as the truth that there are no angels, or that there is no epiphenomenal ectoplasm. For similar reasons, P does not imply such universally quantified truths as the truth (if it is a truth) that all living beings contain DNA molecules.

This loophole can be closed by conjoining to P a "that's-all" statement T , asserting that our world is a *minimal* world satisfying P . Intuitively, this statement says that our world contains what is implied by P , and *only* what is implied by P . More formally, we can say that world $W1$ *outstrips* world $W2$ if $W1$ contains a qualitative duplicate of $W2$ as a proper part and the reverse is not the case. Then a minimal P -world is a P -world that outstrips no other P -world.[*] It is plausible that no world containing angels is a minimal P -world: for any P -world containing angels, there is an angel-free P -world that it outstrips. So $P \& T$ implies that there are no angels. For similar reasons, $P \& T$ implies the other negative and universally quantified statements mentioned above. So the thesis at issue should be that for the relevant macroscopic truths M , $P \& T$ implies M . (From now on we will abbreviate ' $P \& T$ ' as ' PT ', and similarly for other conjunctions among the four statements discussed in this section.)

*[[It is well-known that (just as in the case of defining intrinsic properties), some notion in this circle has to be taken as primitive: perhaps the notion of minimality, or of outstripping, or of containing a qualitative duplicate. One also might invoke the notion of intrinsic property itself, e.g. to say that world $W1$ outstrips if all the intrinsic properties and relations in $W2$ are correspondingly instantiated in $W1$, and if the reverse is not the case.

There is a complication that arises in characterizing the that's-all claim in terms of minimality where a priori entailment, rather than necessitation, is concerned. The relevant claim must hold that the actual world is minimal among the class of *epistemic* possibilities satisfying P , where an epistemic possibility corresponds intuitively to a maximally specific hypothesis that is not ruled out a priori. On some philosophical views, this might come apart from a definition in terms of minimality among the class of metaphysical possibilities. For example, on a view on which it is necessary but not a priori that God exists, 'God exists' might be entailed by P conjoined with an assertion of metaphysical minimality, but not conjoined with an assertion of epistemic minimality. This can be handled straightforwardly by formalizing the notion of epistemic possibility (e.g. Chalmers forthcoming c). Alternatively, where this sort of case is concerned, one can retain the intuitive characterization of the that's-all statement.]]

(2) *Indexical truths*. As described, the information contained in PT specifies a world objectively. For this reason, it does not imply any indexical truths: truths such as 'I am Australian', or 'life evolved in the past', or 'there is water on this planet'. It also may not imply such truths as 'water is made of H₂O'. If the universe contains an inhabited planet where there is a superficially identical liquid made of XYZ, then the information in PT alone will not allow us to decide whether we live on the H₂O planet or on the XYZ planet, so PT is at least epistemically compatible with the claim that water is made of XYZ.

This loophole can be closed by conjoining to PT some *locating* information (or *indexical* information) I . I can be thought of as a "you are here" marker added to the objective map given by PT . I can consist of the conjunction of any two truths 'I am A ' and 'now is B ', where A is an identifying description of myself (or the subject in question) and B is an identifying description of the current time. (An identifying description is a description such that PT implies that there is a unique individual or time satisfying the description.) As long as PT implies all relevant objective information, then I will enable the subject to "locate" himself or herself, and PTI will imply all indexical truths.[*] So the thesis at issue should be that for the relevant macroscopic truths M , PTI implies M .

*[[Strictly speaking, to handle some cases (e.g. the "Two Tubes" case discussed in Austin 1990), I needs to contain further indexical information, such as information about the referent of certain special demonstratives (see Chalmers forthcoming

a). This subtlety can be passed over for present purposes.]].

(3) *Phenomenal truths*. One of the questions at issue is whether phenomenal truths - truths concerning states of phenomenal consciousness - are implied by microphysical truths. Many hold that they are not, and hold that adding indexical and that's-all information does not close the gap. We will not adjudicate that issue here. But we note that if phenomenal truths are not implied by *PTI*, then it is likely that many other macroscopic truths are not so implied either. For example, knowing whether an object is red arguably requires knowing whether it is the sort of object that causes a certain sort of color experience, and knowing whether an object is hot arguably requires knowing whether it is the sort of object that causes experiences of heat. If so, then if truths about color experience and heat experience are not implied by *PTI*, truths about color and heat are not implied either.

This loophole can be closed while remaining neutral on the important issues, by conjoining *Q*, the conjunction of all phenomenal truths, to *PTI*. *Q* will specify the phenomenal states and properties instantiated by every subject bearing such states and properties, at every time.[*] That is, for every subject who is phenomenally conscious at a given time, *Q* will specify precisely what it is like to be that subject at that time. We can also stipulate that *Q* includes any fundamental principles that lawfully govern phenomenal states, or that connect phenomenal states with physical states. The thesis at issue is then whether *PQTI* implies the relevant macroscopic truths *M*, which we can now understand to include truths about color and heat, as well as phenomenal truths.

*[[The specification of phenomenal states in *Q* should use terms that express what Chalmers (forthcoming a) calls "pure phenomenal concepts", picking out phenomenal states directly in terms of phenomenal character, as opposed to relational phenomenal concepts, which pick out phenomenal states in terms of their typical external causes or effects. We will be neutral here about just what is involved in phenomenal character.]]

To see that this is a neutral way of posing the issue, note that if *PQTI* implies all such *M* (as we shall argue), then either (i) *PTI* alone implies all such *M*, or (ii) *PQTI* implies all such *M*, but *PTI* does not. In the first case, the original thesis involving *PTI* will be upheld. (In this case, phenomenal truths will themselves be implied by *PTI*.) In the second case, all failures of *PTI* to imply a relevant *M* will be associated with the failure of *PTI* to imply *Q*, in the sense that adding *Q* will close any epistemic gaps. Putting these cases together, the thesis concerning *PQTI* entails the crucial claim that *if* phenomenal truths are not implied by *PTI*, then there is a special epistemic gap in the phenomenal case.

(If we combine these alternatives with the thesis that reductive explanation goes along with a priori entailment, then the first alternative above leads to the view that all the relevant macroscopic truths, including phenomenal truths, are reductively explainable. The second alternative leads to the view that phenomenal truths are not reductively explainable, and that other macroscopic truths are reductively explainable "modulo phenomenology" - that is, that we can reductively explain those aspects of them for which the phenomenal plays no conceptually constitutive role. This would fit nicely with the view (articulated by Nagel 1974 and Searle 1991, among others) that the actual reductive explanation of phenomena such as color and heat has proceeded by explaining their objective aspects while leaving their subjective aspects untouched.)

Note that the addition of Q to the conjunction changes the way that we define T and I . T will now hold that the actual world is a minimal world satisfying PQ , and I will involve identifying descriptions of a subject and a time relative to the information in PQ . (In fact, if P does not imply Q and Q is true, then the original version of PTI is false.) So the crucial thesis is that $PQTI$, understood as the conjunction of P and Q with T and I so understood, implies M , for the relevant M . Note also that the information in I depends on the specification of a subject and a time, so that strictly speaking, $PQTI$ will vary between subjects, and the thesis will be that $PQTI$ implies M for any given subject and time in our community. In what follows, we will assume that an arbitrary subject and time have been selected.

3 A Priori Entailment and Conceptual Analysis

It is sometimes claimed that for $A \rightarrow B$ to be a priori, the terms in B must be definable using the terms of A . On this view, a priori entailment requires definitions, or explicit conceptual analyses: that is, finite expressions in the relevant language that are a priori equivalent to the original terms, yielding counterexample-free analyses of those terms. This is not our view. The falsity of the claim can be seen from the following.

Let G be the conjunction of the statements in the following passage: 'Smith believes with justification that Jones owns a Ford. Smith initially has no beliefs about Brown's whereabouts. Smith forms a belief that Jones owns a Ford or Brown is in Barcelona, based solely on a valid inference from his belief that Jones owns a Ford. Jones does not own a Ford, but as it happens, Brown is in Barcelona.' Let K be the statement 'John does not know that Jones owns a Ford or Brown is in Barcelona'.

It is plausible that $G \rightarrow K$ is a priori. But it is also plausible that there is no explicit analysis of the concept of knowledge using the terms involved in G . If so, a priori entailment does not require explicit analyses of the terms in the consequent using the terms in the antecedent. It is also somewhat plausible that there is no explicit analysis of the concept of knowledge at all. If so, a priori entailment does not require explicit analyses of the terms in the consequent.

The two important features of the case are the apriority of the conditional $G \rightarrow K$ and the absence of explicit analyses of the concept of knowledge. On the first point: once general skepticism about the a priori is set aside, this conditional seems to be a typical example of an a priori truth. Someone who knows that G is true and who possesses the concepts involved in K (in particular the concept of knowledge) is thereby in a position to know that K is true, even if they lack any further relevant empirical information. That is, a grasp of the concept of knowledge (along with a grasp of the other concepts involved) and rational reflection suffices to eliminate the possibility that G is true and K is false.

This conditional plays an essential role in Gettier's argument for the conclusion that knowledge is not justified true belief. Gettier's argument was an a priori argument, in which empirical information played no essential role, and its conclusion is a paradigmatic example of a non-obvious a priori truth. The argument proceeds by presenting the possibility that G holds, and appealing to the reader's concept of knowledge to make the case that if G holds, K holds (and J holds, where J is a corresponding positive

claim about John's justified true belief). Empirical information plays no essential role in justifying belief in this conditional, so the conditional is a priori. The a priori conditional itself plays an essential role in deriving the a priori conclusion.

On the absence of explicit analyses of knowledge: We take it that this is a reasonable conclusion to draw from four decades of failed attempts to produce explicit analyses. Certainly no explicit analysis has met with widespread approval, and proposed analyses are always confronted quickly by plausible counterexamples. Of course it could be that there is a correct explicit analysis that has not yet been produced, or that has been produced but overlooked. But even if so, it seems clear that the a priori entailment from *G* to *K* is not *hostage* to an explicit analysis of knowledge that would support the entailment. Whether or not there is such an analysis, the entailment is a priori all the same.

So a priori entailment does not require explicit analysis. If anything, the moral of the Gettier discussion is that the reverse is often true: explicit analyses are themselves dependent on a priori intuitions concerning specific cases, or equivalently, on a priori intuitions about certain conditionals. The Gettier literature shows repeatedly that explicit analyses are hostage to specific counterexamples, where these counterexamples involve a priori intuitions about hypothetical cases. So a priori conditionals seem to be prior to explicit analyses at least in matters of explicit justification; and in general there is no reason to hold that a priori conditionals need explicit analyses to underwrite them.

It could be argued that while these a priori entailments are not underwritten by explicit analyses, they are underwritten by explicit *sufficient* conditions for knowledge or its absence: for example, the condition that a belief solely based on inference from a false belief is not knowledge. Of course it is trivial that there is a sufficient condition in the vicinity of such an entailment (the antecedent provides one such), so the claim will be interesting only if the complete set of sufficient conditions for knowledge is not indefinite and open-ended. But the Gettier literature suggests precisely that the set of sufficient conditions for knowledge is open-ended in this way; if it were not, we would have a satisfactory explicit analysis. And as before, the a priori entailments are not hostage to the proposed sufficient conditions; if anything, proposed sufficient conditions are hostage to a priori intuitions about specific cases.

Once an essential role for explicit definitions is eschewed, the model of conceptual analysis that emerges is something like the following. When given sufficient information about a hypothetical scenario, subjects are frequently in a position to identify the extension of a given concept, on reflection, under the hypothesis that the scenario in question obtains. Analysis of a concept proceeds at least in part through consideration of a concept's extension within hypothetical scenarios, and noting regularities that emerge. This sort of analysis can reveal that certain features of the world are highly relevant to determining the extension of a concept, and that other features are irrelevant.

What emerges as a result of this process may or may not be an explicit definition, but it will at least give useful information about the features in virtue of which a concept applies to the world. It will usually be the case that one can find complex expressions whose conditions of application approximate those of the original concept to some degree, where one finds increasingly good approximations through increasingly complex expressions. In this way we can elucidate at least many important aspects of how a concept's

extension depends on the world. But in general, there is no reason to suppose that a finite expression yielding a counterexample-free analysis of a concept must result from this process.[*] This pattern, whereby a conditional ability to evaluate a concept's extension yields elucidation of a concept without a finite counterexample-free analysis, is illustrated very clearly in the case of 'knowledge'.

*[[FJ is somewhat more optimistic than DJC about the possibility of satisfactory finite analyses, especially if one recognizes that conceptual analysis can accommodate an element of conceptual revision to clear up confusions in a folk concept. See Jackson, Pettit, and Smith 2000 for some reasons for optimism.]]

The possibility of this sort of analysis is grounded in the following general feature of our concepts. If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension. For example, if a subject possesses the concept 'water',[*] then sufficient information about the distribution, behavior, and appearance of clusters of H₂O molecules enables the subject to know that water is H₂O, to know where water is and is not, and so on. This conditional knowledge requires only possession of the concept and rational reflection, and so requires no further a posteriori knowledge.

*[[We will use quoted expressions for both concepts and terms in this article, and will slide between the levels of thought and language for ease of discussion; nothing turns on this.]]

Of course this claim is trivial if the empirical information in question is allowed to include information directly involving the concept at issue (e.g. 'water'). But as this case and the case of knowledge suggests, the claim is often true even when that sort of information is excluded: a 'water'-free description of the world can enable one to identify the referent of 'water', and a 'knowledge'-free description of the world can enable one to decide whether a given belief is an instance of knowledge. In these cases, we can say that *nontrivially* sufficient information enables identification of a concept's extension. We will not try to give a precise account of what counts as nontrivially sufficient information (presumably one should also exclude near-synonymous expressions, such as 'aqueous' or 'epistemic'), but will leave the notion intuitive. Note that there is no requirement at this point that the information in question be microphysical, or that it be about perceptual evidence, or that it be the information in *PQTI*.

This ability to identify a concept's extension is not restricted to true empirical information about the actual world. If the world had turned out differently, we could still have identified the concept's extension. Correspondingly, we can evaluate the concept's extension given *hypothetical* information about ways the actual world might be. Let us say that a hypothesis is *epistemically possible* (in the broad sense) when it is not ruled out a priori. Let us say that an epistemically possible hypothesis characterizing the total state of the world corresponds to an *epistemic possibility*: intuitively, a specific way the actual world might turn out to be, for all one can know a priori.[*] Then sufficient information about an epistemic possibility enables a subject to know what a concept's extension will be, under the hypothesis that the epistemic possibility in question is actual. For example, in the Gettier case, it is irrelevant whether or not John's case is actual: a subject can know that *if* John's case as described is actual, then John does not know that someone owns a Ford. Or in the case of water, given appropriate information about the distribution, behavior, and appearance of clusters of XYZ molecules (information analogous to the information we

have about H₂O in the actual world), a subject is in a position to conclude that *if* the information is correct, then water is XYZ.

*[[See Chalmers (forthcoming c) for a formalization of this notion. The informal characterization suffices for present purposes.]]

If something like this is right, then possession of a concept such as 'knowledge' or 'water' bestows a *conditional ability* to identify the concept's extension under a hypothetical epistemic possibility, given sufficient information about that epistemic possibility and sufficient reasoning. That is, possession of these concepts in a sufficiently rational subject bestows an ability to evaluate certain conditionals of the form $E \rightarrow C$, where E contains sufficient information about an epistemic possibility and where C is a statement using the concept and characterizing its extension, for arbitrary epistemic possibilities.[*] And conceptual analysis often proceeds precisely by evaluating conditionals like these.

*[[Determinate application of these conditionals may be restricted to epistemic possibilities that are not too far from home. When epistemic possibilities deviate greatly from our presuppositions about the actual world, some of our concepts will lose determinate application.]]

In the most important cases, these conditionals will be a priori. Certainly there will be related cases in which $E \rightarrow C$ is a posteriori: for example, it is a posteriori that if a glass contains H₂O, it contains water. But these will be cases in which the antecedent E does not contain sufficient empirical information to identify the concept's extension given possession of the concept alone. The a posteriority of these conditionals reflects the fact that further empirical information is required for their justification. But then all we need to do is conjoin E with the relevant further empirical information F , and we will obtain a conditional $E' \rightarrow C$ that is knowable a priori, where E' is the conjunction of E and F . For example, in the case of 'water', identification of the concept's extension requires a great deal of further information about the distribution, behavior, and appearance of clusters of H₂O molecules in the world. But once this information F is conjoined with the original information E , we obtain a more complex conditional of the form $E' \rightarrow C$ that is plausibly a priori.

Given sufficient empirical information in the antecedent, there is good reason to believe that the resulting conditionals will be a priori. These a priori conditionals will reflect the way in which we can identify a concept's referent. If we possess a concept, then sufficient empirical information E enables us to conclusively identify the concept's extension and to know associated truths C , and we are in a position to do this whether the information in E is actual or hypothetical. Because all the relevant empirical information is present in the antecedent of the conditional, empirical information plays no essential role in justifying belief in the conditional. So $E \rightarrow C$ is a priori. We can call this sort of conditional an *application conditional*.

We do not claim that application conditionals with nontrivially sufficient information in the antecedent exist for all concepts. For example, it may be that there is no way to know truths about time without having empirical information that more or less explicitly invokes the concept of time. If so, there will be no antecedents with nontrivially sufficient information to imply consequents about time. Some other

primitive concepts (space? cause? consciousness?) may also be like this. But for many or most concepts, there will exist application conditionals (corresponding to arbitrary epistemic possibilities) whose antecedents contain nontrivially sufficient information.

It should be noted that nothing here conflicts with the conclusions of Kripke (1972) about names and natural kind terms. On Kripke's picture, it is an a posteriori necessity that water is H₂O. But this is entirely compatible with there being an a priori conditional from certain (false) statements about the distribution, behavior, and role of XYZ to 'water is XYZ'. Kripke allows that a conditional can be a priori even when it is not necessary. For example, his view allows that claims such as 'heat (if it exists) is the dominant cause of heat sensations' is a priori, although not necessary. The same goes for many claims with similar form, such as claims about the length of the meter stick in Paris. If so, then many conditionals of the same form as 'if *X* is such-and-such, then *X* is heat' are a priori, even where it is an a posteriori necessity that heat is not *X*.[*]

*[[Some contemporary philosophers (e.g. Salmon 1986) go beyond Kripke's view, holding that statements such as 'heat (if it exists) is the dominant cause of heat sensations' (and the like) are *not* a priori, and that many apparently empirical identities of the form '*X* is *Y*' are a priori. This counterintuitive view is usually held for theory-dependent reasons. For example, if a philosophical theory holds that statements like these express singular propositions (so that '*X* is *Y*' expresses the same singular proposition as '*Y* is *Y*'), and that the apriority of a statement is a function of the proposition it expresses, then given certain assumptions about the apriority of singular propositions the theory will be committed to these counterintuitive claims. We think that this is itself a good reason to reject the theories in question, and that there are also strong independent reasons to reject such theories, but we cannot go into that matter here. For present purposes, we will simply assume that the intuitive claims about apriority are correct.]]

One can put this by making the familiar observation that even if it is not metaphysically possible that water is XYZ, it is epistemically possible that water is XYZ, in the broad sense that it is not ruled out a priori that water is XYZ. It is also epistemically possible that XYZ has a certain specific distribution, behavior, and appearance (of a sort that is characteristic of water). A subject possessing the concept of 'water' can reason straightforwardly that *if* the second hypothesis obtains (in the actual world), then the first hypothesis obtains. This sort of a priori relationship among epistemic possibilities is entirely compatible with different necessary relationships among metaphysical possibilities.

There is also nothing here that contravenes Kripke's epistemological arguments against certain descriptive views of reference-fixing. Indeed, these arguments can be seen as invoking just this sort of reasoning about epistemic possibilities. In considering whether the term 'Gödel' is a priori equivalent to 'the prover of the incompleteness of arithmetic', Kripke considers a certain epistemically possible hypothesis on which the incompleteness of arithmetic was proved by a man named 'Schmidt' and was then stolen and promulgated by a friend named 'Gödel'. Kripke suggests that if this hypothesis is actual, then our term 'Gödel' refers not to Schmidt, who proved incompleteness, but to his friend. If so, it is not a priori that Gödel (if he existed) proved the incompleteness of arithmetic. This is a straightforward example of armchair reasoning about how the extension of a concept depends on how the actual world turns out.

One observation must be made in the case of names. It is possible that two people who use a given name might use it with *different* a priori application conditionals. For example, if Leverrier uses 'Neptune' as a

name for whatever planet perturbs the orbit of Uranus, then conditionals of the form 'If X is the planet that perturbs the orbit of Uranus, X is Neptune' will be knowable a priori by him. But if Leverrier's wife uses 'Neptune' knowing only that it is an astronomical object for which her husband is searching, then the conditional just mentioned will not be knowable a priori by her. She will still have an ability to know the term's extension given sufficient information about the actual world, but for her, the antecedent information will crucially require information about her husband's intentions. So possession of 'Neptune' gives the wife and the husband conditional abilities to apply the term that are somewhat different from each other: they yield different application conditionals, and they yield a different pattern of application across epistemic possibilities.

This sort of epistemic variability suggests that at least in the case of names, we should take the apriority of a sentence to be subject-relative. For example, we can say that a sentence S is a priori for a speaker when the belief that S would express for the speaker can be justified independently of experience, yielding a priori knowledge.[*] So above, 'Neptune (if it exists) perturbs the orbit of Uranus' is a priori for Leverrier but not for his wife. The relevant class of a priori conditionals will then be subject-relative in a similar way. But it remains the case that when a subject possesses a name, the subject will have a conditional ability to identify its extension given sufficient empirical information about the actual world, and the relevant conditionals will be a priori for the subject.

*[[Or we might say: a sentence S is a priori for a subject when an assertion of S by that subject would express an a priori justifiable thought, where thoughts are the sort of propositional attitude tokens apt to be directly expressed by an assertion, including beliefs and belief-like attitudes such as entertainings. The move from beliefs to thoughts accommodates the possibility that the subject does not believe what he or she asserts. An a priori justifiable thought is a thought that can be justified independently of experience, on idealized rational reflection, yielding a priori knowledge.]]

The case of names is incidental to our central purposes, but it is arguable that something similar can occur for a natural kind term such as 'water'. The most obvious cases of this will occur when a subject uses the term with deference to others in a linguistic community. Such cases will be not unlike the case of Leverrier's wife above. For present purposes, it is probably best to take it that we are stipulating that the terms in question are used nondeferentially. Deference raises issues that are orthogonal to the central issues here about reductive explanation here. Nothing important to those issues would change if we were dealing with a community of one speaker, or with only "expert" subjects.

It may be that a natural-kind term can be epistemically variable for reasons independent of deference. For example, perhaps a city-dweller might use 'water' nondeferentially for the liquid that comes out of faucets (knowing nothing of oceans), and a beach-dweller might use 'water' nondeferentially for the liquid in the oceans (knowing nothing of faucets). This sort of case can be treated as we suggested that cases involving names be treated: the subjects have different conditional abilities, and different associated conditionals will be knowable a priori for each of them. But as before, this epistemic variability gives no reasons to deny that the a priori conditionals exist. In what follows, we will usually abstract away from this sort of cross-subject variability, as the central issues about reductive explanation are largely orthogonal to these questions about variable use within a community. But when necessary, all of the claims about apriority in the following can be put in subject-relative terms.

4 On the Entailment of Macroscopic Truths

We can now address the crucial question: for the relevant macroscopic truths M , does $PQTI$ imply M ?

We have already made the case that for many such truths M , there is *some* nontrivially sufficient empirical information E such that E implies M . The question now is whether the specific empirical truth $PQTI$ is such an E ; that is, whether $PQTI$ contains sufficient information to imply M .

A conclusive argument for this thesis would require very detailed discussion, but here we will present some reasons for finding the thesis plausible. The basic argument has a two-step structure. First, $PQTI$ implies complete information (in the language of physics) about the structure, dynamics, composition, and distribution of macroscopic systems, as well as about the actual and potential perceptual appearances that they present. Second, this information about macroscopic structure, dynamics, composition, distribution, and appearance (along with residual information from $PQTI$) implies ordinary macroscopic truths such as M . But we can take things a little more slowly. To address the question, it is useful to start by imagining that one starts with *only* the empirical information specified by $PQTI$, and by asking: could one thereby come to know the truth of M ?

To start with, one can get some distance using Q alone. On a phenomenalist view, Q alone (construed as the complete truth about actual and counterfactual experiences) implies all truths. Even if phenomenalism is rejected, Q still gives a significant epistemic foothold on the world. Combining the complete phenomenal information in Q with the indexical information in I puts me in a position to determine the phenomenal character of my current experiences, and of my experiences throughout my lifetime. This will include in particular the phenomenal character of a lifetime of *perceptual* experiences. This information serves as at least an epistemic guide to many macroscopic truths, just as it does in ordinary life. If V is a specific phenomenal character of a visual experience as of a large object in front of me, then if I know that I am now having an experience with phenomenal character V , then I might reasonably infer that there is a large object in front of me.[*] The same goes for many other perceptual experiences.

*[[Note that as before, Q will characterize the experience according to its phenomenal character alone, leaving causal relationships to the environment unspecified. So Q will not characterize V as the phenomenal character of visual experiences caused by large objects, and knowledge of Q will not build in this sort of causal knowledge. Rather, knowledge of Q will build in knowledge of phenomenal character: that is, knowledge of what it is like to have the relevant experiences. The relevant point here is that knowledge of phenomenal character alone plausibly gives some epistemic guidance about the nature of the environment, even if it leaves open many skeptical hypotheses.]]

Of course this information does not remotely suffice to imply all the relevant macroscopic truths M . First, there is a vast class of truths about which my perceptual experience give no guidance: truths concerning unperceived objects, for example. Second, for those truths M about which my perceptual experience gives guidance, it remains epistemically possible (in the broad sense) that I have these perceptual states but that M is false. Such epistemic possibilities range from traditional skeptical scenarios concerning the nonexistence of the external world to a wide range of scenarios involving perceptual illusions, false

testimony, false abductive or inductive inferences, and so on.

Nevertheless, truths about which my perceptual experience gives no guidance can often be settled by further information; and skeptical scenarios can often be ruled out by further information. Starting with the further information in *Q*, information about others' experiences and about counterfactual experiences will give guidance about many unperceived objects. Of course numerous skeptical scenarios will still be left open, at least if phenomenalism and related views are rejected. But once the information in *P* (and *T* and *I*) is admitted, many such scenarios are immediately ruled out: those in which the physical world does not exist, for example. The residual question is whether the information in *PQTI* suffices to derive knowledge of the unknown truths, and to rule out all the skeptical scenarios.

The information in *P* will play a crucial role. This includes complete information about the structure and dynamics of the world at the microphysical level: in particular, it includes or implies the complete truth about the spatiotemporal position, velocity, and mass of microphysical entities. This information suffices in turn to imply information about the structure and dynamics of the world at the macroscopic level, at least insofar as this structure and dynamics can be captured in terms of spatiotemporal structure (position, velocity, shape, etc) and mass distribution. For example, for any given region of space at a time, the information in *P* implies information about the mass density in the region, the mass density in various subregions, the causal connections among various complex configurations of matter in the region, and the extent to which the matter in the region behaves or is disposed to behave as a coherent system. This information suffices to determine which regions are occupied wholly by causally integrated systems that are disposed to behave coherently. So the information plausibly suffices for at least a geometric characterization - in terms of shape, position, mass, composition, and dynamics - of systems in the macroscopic world.

The central point here is that a macroscopic description of the world in the language of physics is implied by a microscopic description of the world in the language of physics. Such a thesis is extremely plausible: it is not subject to any worries about translation between vocabularies, and involves only a change of scale. The only worry might concern the status of bridging principles within physical vocabulary: for example, is it a priori that the mass of a complex system is the sum of the masses of its parts? If there are any concerns here, however, they can be bypassed by stipulating that the relevant physical principles are built into *P*.[*] *P* also implies information about systems' microstructural composition, and about their distribution of systems across space and time, including the relations between systems (characterized in microphysical terms) and about any given system's history (characterized in macrophysical terms).

*[[In fact, for present purposes one could even stipulate that *P* contains a complete description of the world, both microscopic and macroscopic, in the language of physics. The central issues about consciousness, and the central issues that divide us from Block and Stalnaker, arise equally whether one starts with microphysics or macrophysics, and so are unaffected by such a stipulation. Still, on our view such a stipulation is unnecessary.]]

Further, the information in *P* and *Q* together will imply truths about regularities connecting the physical and phenomenal domains. *PQTI* will include or imply all truths about lawful covariation between the world's physical and phenomenal states. If my own phenomenal states depend directly on associated

physical states, then P and Q in conjunction with I will imply information about the dependence. If there are certain regularities by which other physical systems in the world indirectly affect my perceptual phenomenal states, then P , Q , and I will imply information about those regularities. So although information about the external causes of perceptual phenomenal states is not built in to Q , $PQTI$ will imply information about these causes. It will also imply information about the perceptual phenomenal states that various external systems are disposed to cause when appropriately situated: that is, about the perceptual appearance of these systems.

Overall, $PQTI$ implies complete information about the (geometrically characterized) structure and dynamics of macroscopic systems and objects in the world, their spatiotemporal distribution and microstructural composition, and their actual and potential perceptual appearances. This information puts a subject in a position to conclusively know (on rational reflection) the truth or otherwise of any ordinary macroscopic claim M . Complete knowledge of perceptual appearances yields the information that members of our community rely on in coming to know macroscopic truths; and complete structural, dynamic, distributional, and compositional information contains all the information that we need to settle the truth of claims that perceptual information does not settle.

For example, knowledge of the appearance, behavior, and composition of a certain body of matter in one's environment, along with complete knowledge of the appearance, behavior, and composition of other bodies of matter in the environment, and knowledge of their relationships to oneself, puts one in a position to know (on rational reflection) whether or not the original system is a body of water. The same goes for knowledge of whether or not the system is gold, whether or not it is alive, whether or not it boils at a certain temperature, or whether or not it is found in the oceans. And the same applies to ordinary macroscopic truths M in general: complete knowledge of structure, dynamics, composition, distribution, and appearance puts one in a position to know whether or not M is true.

Further, the information in $PQTI$ serves to conclusively eliminate arbitrary skeptical hypotheses under which M might be false. Hypotheses involving perceptual illusions or hallucinations are eliminated by full structural and dynamical information. Hypotheses concerning differences in the past and the future are eliminated by full distributional information. Hypotheses concerning differences in underlying causal or compositional structure are eliminated by full compositional information. Even skeptical hypotheses concerning differences in others' minds are plausibly eliminated by full phenomenal information. Further skeptical hypotheses about the subject's own relation to these systems, or about their exhaustiveness, are removed by the indexical and that's-all information in $PQTI$. A relevant skeptical hypothesis would have to be one in which the structure, dynamics, distribution, composition, and appearance of objects and systems across space and time is preserved (along with indexical and that's-all information), but on which M is false. There do not seem to be any such: the knowledge outlined above suffices for conclusive knowledge of M .

Importantly, no other empirical information is required to justify the inference from $PQTI$ to M : $PQTI$ contains all the information that is needed to know M . We can imagine a subject engaged in a Cartesian suspension of all empirical belief, and then given the information that $PQTI$. Given this information alone, the subject would be in a position to reconstruct knowledge of the structure, dynamics, distribution,

composition, and appearance of external systems by the reasoning above, and from there to ascertain the truth of macroscopic claims such as M and to eliminate any relevant skeptical hypotheses. All that is required here is possession of the concepts in M , the information in $PQTI$, and rational reflection.

So knowledge of $PQTI$ suffices in principle for conclusive knowledge of M , with no other empirical information required. The same reasoning applies hypothetically: so even without knowing $PQTI$, a subject is in a position to know that if $PQTI$ is true, then M is true, and to rule out any hypothesis on which $PQTI$ is true but M is not. So the subject is in a position to know the truth of the material conditional $PQTI \rightarrow M$.[*] As before, empirical information plays no essential role in justifying knowledge of this conditional (all the information required is present in the antecedent), so the subject is in a position to know the conditional a priori.[*] It follows that $PQTI$ implies M .

*[[One might worry: if M involves a natural kind term, might belief in truths involving the term (even conditional truths) require acquaintance with the relevant kind? In response: even if acquaintance is required for possession of the concept and so for the relevant belief, it does not enter essentially into the belief's justification, for the reasons above. Whatever one says about acquaintance, it is plausible that a subject competent with the terms can in principle use ' $PQTI \rightarrow M$ ' to express a belief, such that the belief is true, justified, and constitutes knowledge, and such that empirical information does not enter essentially into the justification of this knowledge. So $PQTI \rightarrow M$ will be a priori for the subject.]]

*[[A more detailed argument against empirical justification is given in section 5, part (6).]]

As before, a priori knowledge of $PQTI \rightarrow M$ does not rely on any explicit analysis of the concepts involved in M , or on any explicit bridging principles connecting the vocabulary of $PQTI$ with the vocabulary of M . Just as a 'knowledge'-free description of a Gettier situation implies relevant claims about knowledge without requiring an explicit bridge between the vocabularies, $PQTI$ implies the truth of M without requiring an explicit bridge between the vocabularies. Rather, $PQTI$ serves as sufficient information for determining the truth of M , in the sense described earlier. In effect, this breaks down into two stages: $PQTI$ serves as sufficient information for determining complete information about structure, dynamics, distribution, composition, and appearance, and this information serves in turn as sufficient information for determining the truth of M . So $PQTI \rightarrow M$ is an a priori application conditional.

It might be objected that we do not yet have a completed physics, so that we do not yet know what P says, so we cannot know whether $PQTI$ implies M .[*] But for present purposes, we do not need to know what P says. All we have needed here is that P implies truths about the structure and dynamics of macroscopic objects in spatiotemporal terms, along with truths about mass and about microphysical composition. And that implication is likely to be robust over physical theories. (If a physical theory does not yield this implication, we are unlikely to count it as a complete physical theory.) In any case, we stipulated earlier that we are assuming a physical theory that implies truths about the position, velocity, and mass of all microscopic entities. Given this assumption, the step to an analogous characterization of macroscopic systems proceeds as outlined above.

*[[Byrne (1998) makes this sort of objection against an argument for a priori entailment in Chalmers (1996). Byrne's objections in this paper pass over the two-step character of the entailment, from microphysical structure and dynamics to

macrophysical structure and dynamics, and from there (in combination with perceptual appearances) to the relevant macroscopic truths. (This two-step structure is present in Chalmers 1996, albeit briefly.) His discussion also passes over the role of epistemological considerations regarding the elimination of skeptical scenarios.]]

It might also be objected that no human could grasp all the information in *PQTI*, so that no human could grasp the truth of the relevant conditional. This is surely true, but it is no bar to the apriority of the conditional. Apriority concerns what is knowable in principle, not in practice, and in assessing apriority, we idealize away from contingent cognitive limitations concerning memory, attention, reasoning, and the like. Once we idealize away from human memory and processing limitations, the problem here is removed.

We can summarize the argument in a way that makes clear the role of the idealization, by appeal to a fanciful thought-experiment. Imagine a human augmented by a "virtual world" machine. This is a machine containing (i) a supercomputer to store the physical information in *P* and to make the a priori calculations required to move from microscopic structure to macroscopic structure, (ii) a virtual reality device to produce direct knowledge of the phenomenal states described in *Q*, and (iii) tools that use these devices to focus on arbitrary regions of the world, and to deliver information about the macroscopic structure, dynamics, composition, and perceptual appearance of systems in those regions. Using such a machine, a human with no other empirical information could straightforwardly ascertain the truth of the relevant claims *M*. The virtual world does no more than give access to the information contained in *PQTI* and process this information on a priori grounds. So if a human using such a device can ascertain the truth of *M*, it is plausible that *PQTI* $\rightarrow$ *M* is a priori.

One might ask: given that *PQTI* implies ordinary macroscopic truths, does similar reasoning suggest that *PTI* alone implies many such truths? Such reasoning may work for truths concerning the objective spatiotemporal structure of macroscopic systems. But for many or most macroscopic concepts (most obviously concepts such as 'green' and 'hot', but plausibly also concepts such as 'water' and 'tiger'), our application of the concepts relies essentially on associated perceptual appearances. For a typical truth *M* involving such a concept, *PTI* will imply *M* only if *PTI* implies truths about the associated perceptual appearances. And this will plausibly require that *PTI* implies the relevant phenomenal truths in *Q*.

One can then ask: could reasoning such as the above establish that *PTI* implies the truths in *Q*? It seems not. There are familiar reasons (e.g. Jackson 1982) to hold that information about the structure, dynamics, and composition of physical systems does not suffice for information about the character of conscious experiences. And there are familiar reasons (e.g. Chalmers 1996) to hold that there are conceptually coherent alternative possibilities on things are physically just as in *PTI*, but in which different sorts of experiences are present, or in which there are no experiences at all.[*] Whether or not these reasons are sound, their existence makes it clear that there is no direct argument that *PTI* implies *Q* analogous to the argument concerning *PQTI* above. Any argument that *PTI* implies *Q* must rest on quite different considerations.[*]

*[[A reviewer asks whether this be squared with knowledge of other minds. Yes: just as the information in *Q* can be an epistemic guide to truths about the external world without implying those truths, the information in *P* (and information

about one's own mind) might be an epistemic guide to truths about other minds without implying those truths.]]

*[[FJ articulates considerations of this sort in forthcoming work.]]

Note that we have not argued here that *PQTI* implies every truth in every domain (although we are inclined to accept this claim). Given what we have said here, it could be that certain truths in special domains - perhaps concerning mathematics, metaphysics, morality, or mentality? - are not implied by *PQTI*.[*] Presumably this will be because these truths are not conclusively knowable even given full information about structure, dynamics, appearance, distribution, composition and so on. Perhaps these truths are knowable only through different means, or perhaps they are not conclusively knowable at all. But we hope we have said enough to make it plausible that ordinary macroscopic truths concerning everyday macroscopic natural phenomena are implied by *PQTI*.

*[[One might worry in particular about mental truths concerning subjects' propositional attitudes. If these truths are not implied by *PQTI*, then it is likely that many other truths will not be: for example, truths involving social concepts (e.g. 'money') or involving terms used deferentially (e.g. 'arthritis', used by a non-expert). For a more general thesis that applies to such truths, one would need to argue that *PQTI* implies these mental truths (as we hold), or to expand *PQTI* to include mental truths of a more general sort. This question does not affect the present discussion, as truths involving ordinary macroscopic concepts of the natural world (such as 'water' or 'gold', used nondeferentially) do not seem to be tied to propositional attitudes in this way.]]

5 Block and Stalnaker on A Priori Entailment

Block and Stalnaker give a number of arguments against a priori entailments from microphysical truths to macroscopic truths. They do not always cast things in these terms: their discussion often suggests that the issue is whether there is an analysis of macroscopic concepts "in microphysical terms", and whether principles that cross the bridge from microphysical to macroscopic are themselves "microphysical claims". This way of putting things is unfortunate, as on a natural reading these theses are implausible and much stronger than what is needed. But it is clear from the context of Block and Stalnaker's discussion that a priori entailment is the real issue, and many of their arguments are naturally seen as arguments against a priori entailments.

(1) *Explicit analyses*. One aspect of the global structure of Block and Stalnaker's article can be seen as follows. In the first part of their article (especially sections 4-9), they argue there are no explicit analyses of typical macroscopic concepts (such as 'water' and 'life') of a sort that could support an a priori entailment from microphysical to macroscopic truths. In the second part of their article (the long section 10), they discuss the use of the two-dimensional semantic framework as an alternative to explicit analyses, and argue that the existence of the framework does not support the a priori entailments in question. The intended upshot is that the existence of these a priori entailments is doubtful, and that reductive explanation does not require a priori entailment.

An argument with this structure cannot successfully make a case against a priori entailments. To see this, it suffices to note that if this sort of argument succeeds, it succeeds equally in making a case against the a

priori Gettier entailments discussed earlier. It is at least as plausible for 'knowledge' as for 'water' and for 'life' that there is no explicit analysis to support the entailments. And the general criticisms of this use of the two-dimensional framework presumably apply equally to its use in the case of 'knowledge'. But nevertheless, the a priori Gettier entailments discussed earlier exist; or at least, it is clear that this sort of argument does little to make a case against them. So by parity, this sort of argument does little to make a case against the a priori entailments we are concerned with.

We can agree with Block and Stalnaker that there are plausibly no precise explicit conceptual analyses of concepts such as 'water' and 'life' (or at least, no precise analyses of a manageable length). But as we made clear earlier (and as is also made clear in Chalmers 1996 and Jackson 1998), such explicit analyses are not required for a priori entailment.[*]

*[[As part of their argument, Block And Stalnaker suggest the most plausible conceptual analyses are restricted to connections within the same "family" of terms (as with the analysis of 'bachelor' as 'unmarried man'), so that analyses of macroscopic concepts in microphysical terms are implausible. Ned Block has suggested in correspondence that this point extends to a priori entailments, so that the only plausible a priori entailments involve connections within a family. The reasons for accepting this thesis are unclear, however. The main evidence for it seems to lie in the fact that the most familiar and uncontroversial a priori entailments involve intrafamily connections. On the view we are advocating, however, this evidence is easily explained by the fact that there are few *short* crossfamily a priori entailments, since crossfamily entailments will usually have detailed and unwieldy antecedents.]]

We can also agree with Block and Stalnaker that the mere existence of the two-dimensional semantic framework does not imply that the a priori entailments in question exist. As Block and Stalnaker say, the two-dimensional framework provides a good way to capture a priori connections when they exist, but the framework alone does not demonstrate specific a priori connections. But the two-dimensional framework plays no essential role in our arguments for a priori entailments.[*] It played no role in the arguments given here. In the arguments given in Chalmers (1996) and Jackson (1998), it plays only a clarifying role, in removing certain confusions that may arise from the presence of a posteriori necessary connections, and in providing a convenient shorthand for discussing the patterns by which a concept applies to the world. The core of the argument for a priori entailments proceeds independently of these semantic details, just as it does in the Gettier case.

*[[For this reason, we will not address Block and Stalnaker's discussion of the two-dimensional framework in this paper. On our use of the two-dimensional framework, the primary intension (or A-intension) of a concept reflects the a priori conditionals that capture the way a concept applies within various epistemic possibilities, in the manner suggested in section 3 of this paper. This understanding of the framework in epistemic terms differs subtly from Block and Stalnaker's understanding in terms of context-dependence; for discussion, see Chalmers (forthcoming d).]]

Insofar as it is captured by this global structure, then, Block and Stalnaker's discussion does not engage the first-order issue of whether the a priori entailments in question exist. But Block and Stalnaker's article also contains numerous arguments that go beyond what is suggested by this structure. We will discuss these in what follows.

(2) *Conceivability*. In a discussion about conceivability, Block and Stalnaker discuss the issue (raised by

Levine 1993) of whether it is conceivable that P holds without W holding, where P is the complete microphysical truth and W is 'water is boiling'. They discuss two senses of conceivability. In the first sense, conceivability is tied closely to a posteriori possibility and necessity, so that $P \& \sim W$ is inconceivable; but this sense of conceivability is clearly irrelevant to the issue of a priori entailment. In the second sense, P without Q is conceivable if Q cannot be deduced from P using logic and conceptual truths. This sense is more clearly relevant to the issue; we would prefer that the notion be cast in the more general terms of a priori reasoning rather than explicit deduction, but this does not matter to Block and Stalnaker's discussion.[*] They say:

*[[Block and Stalnaker's first sense of conceivability is a version of what Chalmers (forthcoming b) calls "secondary conceivability". Their second sense is a version of "primary conceivability", and in particular is close to Chalmers' "ideal primary negative conceivability". Block and Stalnaker note that this notion is a purely negative notion, so that the name "conceivability" is misleading. It is worth noting that there is also a closely related positive notion in the vicinity, which Chalmers calls "ideal primary positive conceivability".]]

Let C be a complete description, in microphysical terms, of a situation in which water (H_2O) is boiling, and let T be a complete theory of physics. Can one deduce from T , supplemented with analytic definitions, that H_2O would boil in circumstances C ? To see that one cannot, suppose that the deduction is taking place on Twin Earth. The stuff they call 'water' is XYZ, and the process they call 'boiling' is a process that superficially resembles boiling, but that involves a different physical process. Just as they would say (truly) 'water is XYZ, and not H_2O (and if there were H_2O , it wouldn't be water),' so they would say, truly, 'If there were H_2O , and it were behaving like that, it wouldn't be boiling.' They could hardly deduce ' H_2O would boil in circumstances C ' if on their meaning of 'boil', H_2O can't boil at all. (We assume that boiling is a natural kind concept. If you don't agree, substitute some other process term that does express a natural kind concept.)

We can agree with Block and Stalnaker that if there are residents of Twin Earth, they might truly say 'If there were H_2O , and it were behaving like that, it wouldn't be boiling'. But this is irrelevant to the issue of whether there is an a priori entailment from ' H_2O is behaving like that' to ' H_2O is boiling'. The first is a subjunctive conditional about an explicitly counterfactual scenario, and as such is relevant to issues concerning necessity, not apriority.

Compare the following: There is plausibly an a priori entailment from something like 'molecular motion is the dominant cause of our heat sensations' ($M1$) to 'molecular motion is heat' ($M2$). But on a Twin Earth, where heat sensations are caused by something very different, residents might truly say 'if molecular motion were the dominant cause of our heat sensations, it wouldn't be heat'. The latter is a subjunctive claim about an explicitly counterfactual scenario, so at best its truth reflects negatively on the *necessity* of an entailment from $M1$ to $M2$. Its truth clearly has no bearing on the apriority of the entailment (which remains plausible for everything said here). The same goes for microphysics and the boiling of water.

(In fact, the thought experiment does not even bear directly on the necessity of the entailment, since presumably the necessity of an utterance in English does not entail the necessity of the homophonic

utterance in Twin English (witness 'water is H₂O'). If anything, the thought experiment bears on whether it is a priori that the entailment is necessary.)

Block and Stalnaker also say:

We don't really need a Twin Earth story to make our point. Consider a person on actual Earth, who does not know the story about how water boils - perhaps she doesn't even know that water is made up of molecules. One presents her with the theory *T*, and a description (in microphysical terms) of a water boiling situation. Can she then deduce that if *T* is true and a situation met conditions *C*, then the H₂O would be boiling? No, since for all she knows the actual situation is like the one on Twin Earth. Perhaps, if she were told, or could figure out, that the theory was actually true of the relevant stuff in her environment, she could then conclude (using her knowledge of the observable behavior of the things in her environment) that H₂O is water, and that the relevant microphysical description is a description of boiling, but the additional information is of course not a priori, and the inference from her experience would be inductive.

This example is simply an illustration of the general phenomenon whereby physical information must be supplemented by locating information ("you are here") to imply relevant conclusions. The subject needs locating information to know that the theory describes her own environment and not a very different environment. So while this example may show that there is no implication from *P* (the complete physical truth) to *W*, it does nothing to show that there is no implication from *PI* (physical truth supplemented by locating information) to *W*. Of course (as is made clear in Chalmers 1996 and Jackson 1998) the latter is the more relevant claim.

(3) *Uniqueness*

In discussing the possibility of conceptual analyses of terms 'water', Block and Stalnaker spend a considerable amount of time discussing the possibility of analyzing this term as something like 'the waterish stuff around here', where 'waterish' abbreviates a cluster of descriptions of familiar macroscopic properties of water (such as clarity, liquidity, drinkability, and so on). They argue that no such analysis can succeed. The argument proceeds by considering various possible cases where there is more than one sort of stuff in the environment that satisfies the relevant descriptions. Given that we hold that explicit analysis is not required for a priori entailment, we could simply set this discussion aside as irrelevant, but there are some points that are worth discussing.

Block and Stalnaker suggest at one point that for an a priori entailment to go through, water must be analyzed as something like 'the *unique* waterish stuff in our environment'. This seems clearly false. First, it seems obviously epistemically possible (in the broad sense) that there could turn out to be two sorts of stuff with the relevant properties filling the oceans and lakes in our environment, in roughly equal distributions. In such a case, it seems clear that we would say that there are two sorts of water (i.e. this epistemically possible hypothesis is an instance of the epistemically possible hypothesis that there are two sorts of water). So any a priori analysis requiring uniqueness would clearly be incorrect.

Second, the falsity of an analysis involving uniqueness does nothing to rule out a priori entailment. The thought experiment above itself involves largely a priori consideration of an epistemically possible hypothesis. A priori consideration of that hypothesis suggested that it counts as an instance of the epistemically possible hypothesis that there are two sorts of water. So the thought experiment is entirely compatible with an a priori entailment from specific details concerning the hypothesis (in 'water'-free terms) to a conclusion involving water, and indeed may even positively suggest the existence of such an entailment.

If one wanted to come up with an explicit analysis that handles these cases directly, one might start simply with 'the watery stuff in our environment', which handles them well (where the phrase is naturally interpreted to cover all instances of watery stuff in our environment). There will be cases where this analysis does not work: perhaps if there is a single dominant watery stuff and small pockets of different stuff, we might count only the dominant stuff as 'water'. We could then move to something like 'the dominant watery stuff in our environment, if there is one; or any watery stuff in our environment, if there is not'. But no doubt this analysis would itself need to be refined to handle more complex cases, just as in the case of knowledge. It is probably easier, then, to give up the aim of producing a perfect explicit analysis, and to content ourselves with the observation that we have an a priori grasp of how our concepts apply to specific epistemic possibilities, when these are described in sufficient detail. Nothing about these cases contravenes this observation.

(4) *Ghost heat*. In the course of their discussion of uniqueness, Block and Stalnaker discuss a hypothetical epistemic possibility in which it turns out that there is "ghost heat". In this epistemic possibility, things are microphysically just as normal (P is true), so the causal role associated with heat is filled by molecular motion. But the causal role is also filled by a nonphysical substance, which we can call 'ghost heat', in a case of causal overdetermination. Block and Stalnaker say that under this epistemic possibility, it turns out that there are two sorts of heat (molecular motion and ghost heat), so that 'heat = molecular motion' turns out to be false. So P does not imply that heat is molecular motion.

We are not certain that this sort of causal overdetermination is coherent, and we are not certain that if it is coherent, then the epistemic possibility in question should be interpreted as one in which there are two kinds of heat. But even if so, the conclusion is simply an instance of the failure of the microphysical facts P to rule out epistemic possibilities in which there are further nonphysical facts. As before, ruling out these epistemic possibilities requires conjoining a "that's-all" statement T , holding that the actual world is a minimal P -world to P . Once T is conjoined to P (yielding PT), the hypothesis involving ghost heat is ruled out, as a world with ghost heat is clearly not a minimal P -world. So the example shows at best that P does not imply H (where H is 'heat is molecular motion'); it does not show that PT does not imply H . And (as is clear in Chalmers 1996 and Jackson 1998), the latter is the more relevant claim.

Block and Stalnaker anticipate this sort of response, and reply:

Recall that Jackson's definition of physicalism recognized the possibility of a world that is a microphysical duplicate of the actual world, while also containing some additional

nonphysical substances and properties. In response, a "nothing but" condition is built into the definition of physicalism by requiring for the truth of physicalism only that minimal physical duplicates of the actual world be duplicates simpliciter. But the need for this qualification presupposes that the "nothing but" condition is not something that is entailed by microphysics itself; that is, it is not a claim of microphysics that our world is a minimal physical duplicate of itself. To take the "nothing but" condition to be an implicit claim of microphysics would be to build the thesis of physicalism into microphysics, which philosophers such as Jackson and Chalmers, who reject physicalism, should be reluctant to do. They reject physicalism, not microphysics. The truth of what physicists write in textbooks does not depend on the mind-body problem. Even if it is a microphysical fact that H₂O is a waterish stuff around here, it is not a microphysical fact that it is the waterish stuff around here.

The objection here seems to be that the "that's-all" statement *T* is not itself implied by *P*. This is correct but irrelevant. The thesis is not that *T* is part of microphysics. The theses are rather that (i) *PT* (perhaps in conjunction with *QI*) imply the relevant macroscopic truths; that (ii) this sort of implication is required for reductive explanation; and that (iii) an analogous claim (that *PT* necessitates all truths) is required for physicalism. Nothing that Block and Stalnaker say here gives any reason to reject these claims.

(5) *Methodology and simplicity*. In the background of Block and Stalnaker's specific arguments about uniqueness, there lie some more general worries about the role of a posteriori methodological considerations, such as simplicity, in determining the extension of our terms. These worries are not elaborated at length in Block and Stalnaker's discussion, but they come up in a number of places and in a number of different ways.

For example, in the discussion of ghost heat, Block and Stalnaker suggest that it is by using considerations of simplicity that we rule out the possibility of ghost heat. That seems roughly correct. Our observations are compatible with two hypotheses: *P&T* and *P&G*, where *P* contains the relevant physical truths, *T* is a "that's-all" statement, and *G* is a claim about ghost heat. We do not decide between these hypotheses purely through observation, or through deduction, but through the use of simplicity considerations such as Ockham's razor. Something similar goes on in inductive inference, and in abductive inference, and so on. These considerations play a common role in moving from partial information about the world to more complete information about the world.

All this is quite compatible with our position. It is no part of our position that a posteriori methods are never used in determining our concepts' extensions. They are used all the time: in particular, it is by using a posteriori methods that we form empirical hypotheses about how the world is. The information in *PQTI* itself is knowable only a posteriori; and simplicity considerations play a central role in our coming to know aspects of this information. But nothing here bears on the claim that *given* the information about *PQTI* (which rules out the *P&G* hypothesis), the application of our concepts is determined a priori.

There is one passage that may bear directly on the question above. After considering our intuitions about how the term 'water' might apply to various epistemic possibilities, including possibilities in which there

are multiple substances that play the relevant role, Block and Stalnaker write:

It is a part of the semantics of natural kind terms that they are natural kinds, but it may also be part of the semantics of these terms that this is a defeasible condition. What is not plausibly part of the semantics, something we all know in virtue of knowing our language alone, is what to say in all the myriad cases in which the defeasible condition is defeated. In these cases, what we should say will no doubt be dictated by principles of "simplicity", conservativeness, etc.

Block and Stalnaker do not say anything to back up these claims directly, or to argue against an opponent who holds that these conditions of application to various epistemic possibilities are as much a (tacit) part of the concept of 'water' as the conditions of application brought out in the Gettier literature are a (tacit) part of the concept of 'knowledge'. It should also be noted that even if these conditions of application are not part of the semantics of 'water' in English, this does not entail that a subject's application of the term to epistemic possibilities is not justified a priori. As we saw above in the case of 'Neptune', it may be the case that the relevant conditionals involving a term may vary between users of a term (so that the corresponding conditions of application are not built in to the term's semantics in English), but that each user's knowledge of the conditionals is justified a priori all the same. So Block and Stalnaker need to give substantive arguments against the a priori knowability of the conditionals in question.

Block and Stalnaker go on to discuss the case of 'rheumatism', in which as scientists acquired more empirical informations about various diseases underlying a central syndrome, the term began to be used for the syndrome rather than the diseases. And they discuss the case of 'jade', in which we discovered two underlying substances and decided that there are two sorts of jade. Nothing here contravenes the claim that these decisions manifested a tacit grasp of a priori conditions of application of the terms in question, and that the relevant conditionals (from a full specification of the empirical details to conclusions about jade and rheumatism) are a priori.

Of course the claim should not be made too strong. There is almost certainly a high degree of indeterminacy in our concepts, in their application both to the actual world and to hypothetical epistemic possibilities. It can sometimes happens that when an epistemic possibility is found to be actual, no clear decision about the concept's application is dictated. In such a case there often ensues (explicit or implicit) terminological stipulation or refinement, influenced by various sociological and pragmatic factors. It may be that in the 'rheumatism' case, for example, the initial extension of the concept was indeterminate between applying to the syndrome in general, and applying to the most common underlying disease. If so, there is an equally rational possible history in which the term comes to be applied to the disease; but in our world, it came to be applied to the syndrome.

Such cases can naturally be seen as cases of mild conceptual change, or of conceptual refinement. In these cases, a term that formerly expressed one concept comes to express a slightly different concept, so that (i) the term's extension becomes more determinate, (ii) certain sentences whose truth-values were previously indeterminate come to express truths, and (iii) certain conditionals that were not a priori come to be a priori. But these changes happen in parallel, and there is no time at which a concept's extension is not

determined a priori by truths about the underlying state of the world.

Conceptual change is the exception rather than the rule. Most of the time, decision concerning a concept's extension in response to empirical developments are dictated rationally rather than arbitrarily. Cases of this sort pose no threat to the a priori entailment thesis. To see this, note that in these cases, the relevant empirical developments could be presented in advance as an epistemically possible hypothesis, and decisions concerning the concept's extension under that hypothesis will themselves be rationally accessible in principle. So the empirical knowledge in question plays no essential role in justifying the conditional from the empirical developments to a conclusion about the concept's extension. So in these cases we have no argument against a priori entailment.

(6) *A priori knowledge vs. armchair knowledge.*

A more general worry may be lurking in the background. It may be that Block and Stalnaker are prepared to concede that in many cases, a conditional from a sufficiently complete description of an epistemic possibility to a conclusion about a concept's extension is knowable "from the armchair", but they may deny that it is a priori. This view is suggested by the following passage (p. 43):

This seems to be armchair reasoning, reflection that does not include any obvious reference to real experiments, so it is tempting to conclude that this reflection just unfolds our concepts in a totally a priori way. But what this conclusion misses is that our reasoning about the proper epistemic response in various counterfactual situations is informed not only by our concepts, but by implicit and explicit theories and general methodological principles that we have absorbed through our scientific culture - by everything that the "we" who are performing these thought experiments believe. What people should rationally say in response to various hypothesized discoveries will vary depending on their experience, commitments and epistemic priorities.

On this view, armchair knowledge of application conditionals depends in part on general empirical background knowledge. For example, the empirical background knowledge that the world is simple may play a role in our requiring that the extension of a concept be as simple as possible. Or perhaps the knowledge that the world contains numerous natural kinds may make us more inclined to apply terms in the manner typical of natural kinds. The view can be put as follows. For relevant truths M : (i) a subject can know an application conditional $PQTI \rightarrow M$ from the armchair; (ii) there is some general empirical information E , acquired at some time in the past, that plays a role in justifying the subject's knowledge of the conditional; and (iii) empirical knowledge such as E plays an *essential* role in justifying knowledge of the conditional. If (i) and (ii) are correct, the conditional is not known a priori; if (iii) is also correct, the conditional is not knowable a priori.

We argued in section 4 that empirical knowledge E does not play this sort of justificatory role. But it is useful to further examine the roles that empirical information might play in knowledge of application conditionals. There is no question that empirical information can play a *causal* role in acquiring this knowledge. Empirical knowledge often plays a causal role in the acquisition of concepts with certain a

priori connections, and it sometimes play a role in triggering changes in the a priori connections associated with a term, as in the cases above. There is also no question that E could play a *mediating* role in our knowledge of an application conditional. In some cases, E might itself be implied by $PQTI$, and one could then straightforwardly use E in combination with $PQTI$ to deduce M . But neither of these possibilities entails that E plays an essential role in *justifying* knowledge of the relevant conditionals. We suspect that the cases in which Block and Stalnaker think that empirical information has a justificatory role are in fact cases where the information has a causal or mediating role.

To examine this matter in the context of the current objection, we will take it for granted that considerations earlier in the paper establish that for empirical truths M , the conditional $PQTI \rightarrow M$ is at least knowable from the armchair. The same considerations suggest that the truth or falsity of conditionals of the form $X \rightarrow M$ will be knowable from the armchair, for (false) antecedents X representing specific epistemic possibilities that are analogous to but distinct from $PQTI$. (Such possibilities might include different configurations of physical and phenomenal properties, and possibilities involving different laws of nature.) Further, the earlier considerations suggest that the relevant armchair knowledge is sufficient to justify hypothetical reasoning from the hypothetical acceptance of X to the conclusion that M . (Note that this entails that our knowledge of these conditionals is not essentially grounded, uninterestingly, in armchair knowledge that X is false.) We can call the relevant conditionals *armchair conditionals*. Now the question is: insofar as an empirical claim E plays a role in knowledge of armchair conditionals such as $PQTI \rightarrow M$, does it play a justificatory role or merely a causal or mediating role?

To adjudicate this issue, we can focus on the status of E itself. By the usual reasoning, knowledge of $PQTI$ will enable a subject to know from the armchair whether or not E is true. For example, the information in $PQTI$ puts us in a position to know that the world is reasonably simple.[*] So there will be corresponding armchair conditionals of the form $PQTI \rightarrow E$. Further, since E is a posteriori, there will also be epistemic possibilities (in the broad sense) under which E is *false*: e.g., epistemic possibilities under which the world is not simple. So there will be antecedents X (e.g. describing a non-simple world) such that $X \rightarrow \sim E$ is an armchair conditional. Although the epistemic possibility specified by X falsifies E , we can nevertheless evaluate it using various ordinary concepts, yielding corresponding armchair conditionals $X \rightarrow C$.

*[[An opponent might try to avoid this line of argument by holding the truth or falsity of E cannot be settled even by armchair reasoning from antecedents such as $PQTI$ and X . This position would require a very different sort of argument, however, arguing directly against the considerations in favor of armchair knowledge in section 4. The position is also implausible in Block and Stalnaker's favored case of simplicity.]]

We can then ask: does the role played by E in knowledge of armchair conditionals $PQTI \rightarrow M$ extend to a role in knowledge of armchair conditionals $X \rightarrow C$ where X falsifies E ? (E.g., does our knowledge that the world is simple play a role in our application of concepts to a non-simple epistemic possibility?) If E plays a causal role in knowledge of application conditionals, we would expect the answer to be yes. If E plays a mediating role, we would expect the answer to be no. If E plays a justificatory role, it is not clear what we would expect the answer to be. In fact, one can argue that whether the answer is yes or no, E plays no essential role in justifying the application conditionals.

If the answer is yes: then for such an X and C , $X \rightarrow C$ and $X \rightarrow \sim E$ are armchair conditionals, so the subject can reason hypothetically from the information that X to the conclusion that C and $\sim E$. So if the subject were to discover that X were actual (assuming rationality), the subject would come to know that C and $\sim E$. E would then play no essential role in justifying the subject's knowledge that C , as the justification for items of knowledge cannot be essentially grounded in a falsehood. A subject who rationally evaluates X as a hypothetical can come to conclude C (under that hypothesis) by the same rational process; so E plays no essential role in justifying acceptance of the conditional from X to C . So in this application of the concepts involved in C , E plays no essential justificatory role. At most, E plays a causal role in the subject's acquiring a concept with this pattern of application.

If the answer is no: then the justificatory role of E extends at most to armchair conditionals $Y \rightarrow C$ such that there is armchair justification for $Y \rightarrow E$. Further: if there is armchair justification for $Y \rightarrow E$, then if E justifies $Y \rightarrow C$, $Y \rightarrow E$ will also justify $Y \rightarrow C$ without antecedently assuming E (by hypothetical reasoning from Y to E to C). So justification of $Y \rightarrow E$ is required for E -involving knowledge of $Y \rightarrow C$, and given this justification, E plays no further essential role in justifying $Y \rightarrow C$. In particular, justification of $PQTI \rightarrow E$ is required for E -involving knowledge of $PQTI \rightarrow C$, and given this justification, E plays no further essential role in justifying $PQTI \rightarrow C$. This suggests strongly that in these cases, E plays only a mediating role in knowledge of $PQTI \rightarrow C$.

The only residual question about this second class of cases is what justifies the armchair conditional $PQTI \rightarrow E$. Knowledge of this conditional is not essentially justified by E itself. To see this, note that a subject who is given the information in $PQTI$ and who antecedently suspends judgment about E will nevertheless be in a position to conclude that E holds. Further, the same sort of factors are plausibly relevant in justifying $PQTI \rightarrow E$ and in justifying $X \rightarrow \sim E$ for those X that falsify E ; but E cannot justify the latter, for the reasons given above. Alternatively, $PQTI \rightarrow E$ might be justified by some other empirical factor F ; but then knowledge of E does not play an essential role in justifying $PQTI \rightarrow C$, contrary to the original hypothesis. Knowledge that F is all that is required; and we can run the same argument for F . The same goes for the disjunction of E and F , and indeed for any empirical claim. So no empirical claim plays an essential role in justifying knowledge of armchair conditionals of the form $PQTI \rightarrow C$. At most, such claims play a mediating role.

To illustrate the situation, we can take the case of simplicity, and the example of $PQTI \rightarrow M$. On both alternatives described here, a posteriori simplicity considerations might play some role in adjudicating the status of M . On the first alternative, this is because these considerations have played a causal role in the subject's possessing a concept whose a priori conditions of application across all epistemic possibilities involve simplicity. On the second alternative, it is because the information that the world is simple is itself derivable from $PQTI$, and thereby plays a role in the a priori process of reasoning from antecedent to consequent. Either way, the a posteriori considerations play no role in justifying the conditional itself, and are therefore no bar to its apriority.

It might be objected that if empirical considerations can play a role in affecting what is a priori involving a concept (as in the first alternative above regarding simplicity, and also on the version of the

'rheumatism' case involving terminological stipulation), then the notion of apriority is being watered down. But this seems wrong: apriority is a matter of non-empirical *justification*. Concept acquisition is usually empirically driven, and conceptual drift can occur in response to empirical factors, but neither of these is any bar to the apriority of resulting claims involving the concepts.

A related objection notes that if the apriority of various sentences can be affected by empirical developments, then there is no guarantee that claims we now regard as a priori will turn out to be a priori; as a result, any metaphysical and explanatory conclusions that rest on claims about apriority are suspect. But this is wrong. Even where the sort of conceptual drift described here occurs, the relevant claims at the first stage are still a priori; it is just that later, the same sentence is used to express something different that is not a priori. The change here is merely terminological. Whether people use 'rheumatism' for the syndrome or the disease, the substantive issues concerning the phenomenon itself are the same; and whether people choose to use terms in a way that requires simplicity or not, the substantive issues concerning the phenomena are the same. The change in language simply means that we express things differently.

It may even be that in the future, people may come to use a term such as 'consciousness' or 'life' with a priori application conditions that differ from ours, due to sociological or pragmatic factors, or terminological stipulation, or terminological drift. But this sort of future terminological change has no bearing on the truth or the apriority of claims that we currently make using the term, and has no bearing on any metaphysical or explanatory conclusions that might follow.[*] For example, it could turn out that due to this sort of drift, what someone later calls 'consciousness' can be reductively explained; but that does not imply that consciousness can be reductively explained.

*[[This objection reflects a point that occasionally arises in discussion of the two-dimensional framework. Cannot the primary intension (or A-intension) of a term change, and if so, does this not call any conclusions drawn from a priori reasoning into question? In response: yes, the primary intension of a term can change (though this sort of change is not especially common and is usually minor); and no, this terminological change has no bearing on conclusions drawn from a priori reasoning with our current concept. A related but better objection holds that we can be wrong about the primary intension of a concept because we have not reasoned sufficiently deeply about what to say about a given epistemic possibility. This can certainly happen, but it does not cast doubt on the a priori method. Rather, it suggests that the a priori method has to be practiced well.]]

We conclude that none of Block and Stalnaker's arguments against a priori entailment are compelling.[*]

*[[Yablo (forthcoming) gives some further arguments against the case for a priori entailment, based partly on the role of "sensibility" in deriving macroscopic knowledge from microscopic knowledge. Chalmers (forthcoming b) replies to these arguments.]]

6 Does Reductive Explanation Require A Priori Entailment?

Block and Stalnaker hold that reductive explanation does not require a priori entailment.[*] In arguing for this thesis, their main strategy involves arguing that in paradigm cases of reductive explanation, no a

priori entailment exists. We have already rebutted these arguments. But they also offer a positive model of reductive explanation with some arguments in support. So it is worth saying something here about what reductive explanation involves. We start by examining a clear way in which reductive explanation can succeed and a clear way in which it can fail.

*[[A close tie between reductive explanation and a priori entailment is suggested by Chalmers 1996, Levine 1993, and Loar 1997.]]

When a concept of some natural phenomenon supports a priori entailments from the microphysical, there is a clear sense in which the phenomenon can be reductively explained. These a priori entailments might not support a *reduction* of the phenomenon in question to a microphysical phenomenon (at least in some senses of this term), perhaps because such entailments are compatible with multiple realizability. But nevertheless, in showing how any instance of the phenomenon is itself implied by microphysical phenomena, we show that there is a sort of transparent epistemic connection between the microphysical and macroscopic phenomena. Both the microphysical and the macroscopic phenomena are epistemically contingent, in that they involve the actualization of just one of a host of coherent epistemic possibilities. But where this sort of transparent entailment is present, the epistemic contingency in the macroscopic phenomena is reduced to the epistemic contingency in the microphysical phenomena: there is no further epistemic contingency in the connection.

When a phenomenon is entailed a priori by *PQTI* rather than by just *P*, something similar applies. To be sure, there may be some further epistemic contingency in *T*, *I*, and possibly *Q*, so that these truths themselves may not be reductively explained. But we generally count reductive explanation "modulo totality, indexicality, and consciousness" - that is, reduction of the epistemic contingency of a phenomenon to the epistemic contingency in *PQTI* - to be reductive explanation enough. *T* and *I* are sufficiently minor additions to the reduction base that they do not change much; and while *Q* is a larger addition, scientists are prepared to put the explanation of the subjective aspects of a phenomenon (such as color and heat) on hold and settle for an explanation of the objective aspects. This "carving off" strategy arguably does not yield a complete reductive explanation of these phenomena; but at least we know just what we are not explaining.

So in a reductive explanation of a phenomenon such as water or life, we find that a low-level account of the physical processes involved will in principle imply and explain truths about the macroscopic structure, dynamics, behavior, and (in conjunction with *Q*) appearance of relevant systems. And our concepts of 'water' or 'life' dictate that systems with appropriate sorts of structure, dynamics, behavior, and appearance automatically qualify as water or as alive (at least if they are appropriately situated in our environment, or are relevantly related to systems in our environment). So the relevant microphysical truths (perhaps in conjunction with *Q*, *T*, and *I*) imply the existence of water or of life, and their existence is reductively explained. The same applies to various specific features of water and of life, which can be implied and explained in similar ways.

The corresponding view on the problem of consciousness is what Chalmers (1996; 1999) has called "type-A materialism": the view that there is an a priori entailment from *PTI* to the phenomenal truths *Q* (perhaps

because phenomenal concepts are functional concepts). It is clear that *if* type-A materialism is true, then the phenomenal can be reductively explained in terms of the physical, for much the same reason as above. In this case, there will be a transparent epistemic connection from the physical to the phenomenal, and any epistemic contingency in the phenomenal truths will be reduced to epistemic contingency in physical truths. (And consequently, all those other macroscopic truths that are reductively explainable "modulo consciousness" will be reductively explainable *simpliciter*.) As above, a priori entailment supports reductive explanation.

Next consider a clear case of a failure of reductive explanation. Let us say that the property dualist is right about consciousness, and that consciousness is connected to the physical only by contingent laws, perhaps ultimately by fundamental psychophysical laws. These laws might be inferred, in principle, from psychophysical regularities in the actual world. Given the presence of these laws, we can still arguably have some sort of explanation of consciousness and its properties, in terms of physical processes and the psychophysical laws. From *PTI* plus psychophysical laws, the various specific phenomena of consciousness will be implied. But this will not be a case of *reductive* explanation, precisely because of the need for principles in our explanatory base over and above what is present in *P*. The laws themselves are not explained: they are epistemically primitive, in that they are not implied by more basic truths.[*] And these substantive, epistemically primitive principles play a central role in the explanation of the phenomena. So there is no transparent explanation of the phenomena in physical terms alone, and reductive explanation fails.

*[[That is: they are not implied by truths that do not themselves involve phenomenal states. On a Humean view of laws, fundamental psychophysical laws will be implied by truths about the distribution of physical and phenomenal states; but this sort of implication is irrelevant to reductive explanation, as phenomenal states are what we are aiming to explain. On a non-Humean view of laws, fundamental psychophysical laws will be straightforwardly epistemically primitive.]]

Now let us consider a third sort of case, which corresponds to the way Block and Stalnaker envisage the situation with respect to consciousness, and to the way they envisage reductive explanation in general. Here, as in the case above, we observe numerous psychophysical regularities between brain states and states of consciousness. We are led to infer an underlying *identity* between various brain states and phenomenal states. These identities are not entailed a priori by *PTI*; and correspondingly, the truths about consciousness are not entailed a priori by *PTI*. But the truths about consciousness are entailed a priori by *PTI* plus the identities (at least if truths about brain states are entailed a priori by *PTI*). So we still have a sort of explanation of consciousness and its properties, in terms of physical processes and psychophysical identities.

This third case is a version of what Chalmers calls "type-B materialism", on which physical truths necessitate phenomenal truths without entailing them a priori. We think that this sort of case cannot occur, but we will set that worry aside for the moment, and pretend that it can occur. Assuming that it can occur, is the third case more akin to the first case or to the second case? There are different respects in which it resembles each. Ontologically, it is more akin to the first case, involving type-A materialism, since it is compatible with a materialist ontology on which the explained phenomenon is ultimately physical. But epistemically, it is more akin to the second case, involving property dualism.

Like the second case, the third case yields no transparent explanation of consciousness in terms of physical processes. At best, there is an explanation in terms of physical processes *plus* psychophysical identities. And epistemically, the psychophysical identities play exactly the same role as psychophysical laws. They are inferred from regularities between brain processes and consciousness, in order to systematize and explain those regularities. And most importantly, the identities are not themselves explained, but are epistemically primitive. As with the second case, it is precisely because we need these epistemically primitive psychophysical principles to explain the phenomenon that transparent reductive explanation fails.

Ontologically, these identities may differ from laws. But epistemically, they are just like laws. They are epistemically primitive psychophysical "bridging" principles that are not themselves explained, but that combine with physical truths to explain phenomenal truths. An explanation of the phenomenal will have two epistemically irreducible components: a physical component and a psychophysical component. By calling the bridging principles identities rather than laws, this view may preserve the ontological structure of materialism. But the explanatory structure of this materialist view is just like the explanatory structure of property dualism.

Identities play a role in reductive explanation in other domains, of course. Identities involving heat, or temperature, or genes all have some explanatory work to do. But in these cases, the identities are not epistemically primitive. The identities between heat and the motion of molecules and the identity between genes and DNA, for example, are themselves implied by *PQTI* (at least insofar as they are true), for the reasons canvassed earlier in this paper. So in these cases, both macroscopic truths about the phenomena and the identities themselves are transparently explained.

It is sometimes held that "identities do not need to be explained" (e.g. Papineau 1993). Block and Stalnaker say something similar ("Identities don't have explanations"). But this seems to conflate ontological and epistemological matters. Identities are ontologically primitive, but they are not epistemically primitive. Identities are typically implied by underlying truths that do not involve identities. The identity between genes and DNA, or between water and H₂O, is implied by the underlying truths in *PQTI*, for example. Once a subject knows all the truths about DNA and its role in reproduction and development, for example, the subject will be in a position to deduce that genes are DNA. So this identity is not epistemically primitive.

Of course like other truths involving macroscopic phenomena, subjects do not typically come to know these identities by deducing them from microscopic truths. But the identities are so deducible all the same, and their deducibility is what makes the phenomena in question reductively explainable. If the identities in question were epistemically primitive, then explanations of the macroscopic phenomena in terms of microscopic phenomena would have a primitive "vertical" element, and science would have established a far weaker explanatory connection between the microscopic and the macroscopic than it actually has.[*]

*[[One could plausibly go further, and argue that a scientist's warrant to accept a micro-macro identity in the first place

depends on a warrant to accept in-principle deducibility of the macroscopic phenomena from underlying processes. In the absence of warrant to accept such deducibility, scientists will only be warranted in accepting correlation, not identity.]]

Something similar applies even to the case of "Mark Twain is Samuel Clemens", which Block and Stalnaker discuss. A subject who knows all the qualitative truths in question - physical, mental, social - and who possesses the concepts of 'Mark Twain' and 'Samuel Clemens' will be in a position to deduce that the identity is true, even if the subject is initially ignorant of it. The subject will be in a position to know that there was an individual who was known to his parents as 'Samuel Clemens', who wrote books such as *Huckleberry Finn* and the like under the name of 'Mark Twain', whose deeds were causally responsible for the current discussion involving 'Mark Twain' and involving 'Samuel Clemens', and so on. From all this information, the subject will be able to easily deduce that Mark Twain was Samuel Clemens, and the deduction will be a priori in the sense that it will not rely on any empirical information outside the information specified in the base. So this identity is not epistemically primitive.

We might imagine that there are eliminativists who deny the existence of Mark Twain. They accept the existence of Samuel Clemens, and the description of his exploits: that he went under the name 'Mark Twain', that he published many of the works that we know of under that name, and so on. But they deny that Clemens was Twain: Clemens used the name, but Twain never really existed. The response to such an eliminativist seems clear. Once the eliminativist concedes the full description of Clemens' exploits and of his connection to our current use of the name 'Mark Twain', we should say: if you grant all that, you grant that Clemens was Twain. Once the qualitative description is given, there is no further fact about Twain's existence of which we are ignorant. There is at most a terminological decision to make, and the terminological decision seems clear. So the identity of Twain with Clemens is itself deducible from underlying truths.

It may be that there is a way in which it sounds odd to ask for an explanation of why Mark Twain is Samuel Clemens, or of why water is H₂O. This is partly because we naturally interpret these phrases as asking for a causal or historical explanation. When asked why Mark Twain lived in Missouri, we will give a causal explanation of the factors that led him to be there and to stay there. It is not so easily to causally explain why H₂O is water, or why Samuel Clemens is Mark Twain (although even here, historical explanation might do at least some work).

Setting aside causal/historical explanation, it seems no harder and no easier to *reductively* explain the identities that Mark Twain is Samuel Clemens or that water is H₂O than it is to reductively explain the non-identities that Mark Twain lived in Missouri or that there is water in the Pacific Ocean. The slight sense of linguistic oddness is equally present all these cases, so there is nothing special about identities here. What matters is that in all these cases, the relevant truths are not epistemically primitive: they are implied by various underlying truths, and none of them are needed to deduce truths that cannot be deduced without them. So there is nothing here that gives support to the sort of epistemically primitive micro-macro identities that Block and Stalnaker need.

We have argued that (1) there is no reason to believe that there are epistemically primitive micro-macro identities; and (2) even if there were any epistemically primitive micro-macro identities, they would not

support transparent reductive explanation of the macroscopic phenomena, of the sort found in most cases of reductive explanation.

Of course an opponent could hold that such identities (if they existed) would support "reductive explanation" in *some* sense - e.g., a sense wherein a reductive explanation is simply an explanation that involves an ontological reduction. Nothing rests on the terminological issue, however. The significant points are (1) that this sort of explanation would be very much unlike most paradigmatic cases of reductive explanation, and (2) that by invoking an epistemically primitive bridging structure, the explanatory structure of this theory would be more akin to that of property dualism rather than that of standard materialism. We might say that this sort of theory would be ontologically reductive, but epistemically nonreductive. With this sort of theory, as with a property dualist theory, the explanatory gap between physical and phenomenal might be bridged, but it would not be closed.

7 Does Materialism Require A Priori Entailment?

We have argued elsewhere that if there are phenomenal truths that are not implied by *PTI*, then materialism is false.[*] Block and Stalnaker deny this, endorsing a version of type-B materialism, with a conceptual gap between physical and phenomenal but no ontological gap. They spend much less time on materialism than they do on reductive explanation, however. Their main counterargument again proceeds via the claim that there are no such a priori entailments for most macroscopic phenomena (which are nevertheless clearly physical). We have already addressed these arguments.

*[[Chalmers 1996; Jackson 1994; Jackson 1998.]]

We will not discuss the tenability of type-B materialist views in depth, as there is a large recent literature on the topic.[*] But it may be worthwhile to use what has gone already to outline one way of seeing what we think is wrong with the view.

*[[Balog 1999; Chalmers 1999; Hill and McLaughlin 1999; Loar 1997; Levine 1998; Perry 2000; Yablo 2000; and many others.]]

The standard form of type-B materialism invokes identities to cross the explanatory gap, between phenomenal states and physical or functional states. The identities are epistemically primitive, for reasons discussed above. Of course if we are right about what has gone before, this makes the physical-phenomenal identity unlike any other micro-macro identity, in which there is no analogous epistemic gap. And indeed, it makes the physical-phenomenal identity unlike any other identities, none of which are epistemically primitive in the same way.

A type-B materialist might bite the bullet on these things and hold that psychophysical identities are *sui generis*. In response, one can argue that identities between natural phenomena *cannot* be epistemically primitive. The point where one finds objective (non-indexical) epistemically primitive regularities among natural phenomena is precisely the point at which one finds fundamental natural laws. And one can argue

that what it is to be a fundamental law of nature is precisely to be an objective, epistemically primitive counterfactual-supporting regularity. If this is right, then if there are epistemically primitive psychophysical regularities, they must be regarded as fundamental natural laws.

A more general version of type-B materialism holds that the physical truths *necessitate* the phenomenal truths, whether or not there are any psychophysical identities. One can argue against this view in a similar way. The view requires that there are epistemically primitive necessities connecting natural phenomena; but one can argue that all necessities connecting natural phenomena are implied by *PQTI*. Even Kripkean a posteriori necessities can be seen to be derivable by a priori reasoning from nonmodal empirical truths. So these necessities will be unsupported by analogies with any other clear cases of necessities.

Again, an opponent can bite the bullet and hold that these necessities are *sui generis*.[*] And again, one can respond by arguing that necessities cannot be epistemically primitive in this way, and that the only epistemically primitive necessities connecting natural phenomena are natural (or nomological) necessities.

*[[Among type-B materialists, Balog 1999 and Loar 1997 (and possibly others) accept that psychophysical necessities may be *sui generis* in this way. Balog 1999 argues that reasoning from an epistemic gap to an ontological gap is invalid, since a zombie (a physical duplicate of a conscious being, without consciousness) could engage in similar reasoning with true premises and a false conclusion. Balog's argument crucially requires the premise that a zombie's utterance of 'I am phenomenally conscious' is true. But there are good reasons to hold that such an utterance would be false or truth-valueless. (In an argument between a zombie realist and a zombie eliminativist, the eliminativist would be closer to the truth.) This conclusion is supported on independent grounds in Chalmers (forthcoming a).]]

A more detailed treatment of the arguments here requires invoking the two-dimensional semantic framework. We will not go into those details here.[*] Note that it is only at this point that the two-dimensional framework enters our arguments in an essential way, and not earlier, as Block and Stalnaker suggest.

*[[Very briefly: When the two-dimensional framework is used to argue against type-B materialism, the crucial thesis is that every a posteriori necessity has a metaphysically contingent primary intension. A posteriori necessities of the Kripkean variety all appear to have a metaphysically contingent primary intension, and are therefore compatible with the thesis. But it can straightforwardly be shown that if the thesis is true, type-B materialism is false. The only way out for the type-B materialist is to deny the thesis, accepting "strong necessities" that do not have a metaphysically contingent primary intension. The upshot is that on this view, (1) psychophysical necessities are *sui generis*, and that (2) one is committed to an underlying very strong modal dualism, with two distinct modal primitives (corresponding to spaces of conceptually possible and metaphysically possible worlds). Chalmers (1999) argues that this modal dualism is unacceptable.]]

8 Conclusions

We have argued that ordinary truths about macroscopic natural phenomena are entailed a priori by the combination of physical truths, phenomenal truths, indexical truths, and a that's-all statement. We have argued that reductive explanation requires a priori entailment, and (more briefly) that physicalism requires a priori entailment. If that is right, then if the phenomenal is reductively explainable in terms of the physical, then there is an a priori entailment from physical truths, indexical truths, and a that's-all

statement to phenomenal truths. Contrapositively, if there is no such entailment, then the phenomenal is not reductively explainable in terms of the physical.

Acknowledgments

Thanks to Ned Block and to two *Philosophical Review* referees for comments.

References

- Austin, D.F. 1990. *What's the Meaning of "This"?* Ithaca, NY: Cornell University Press.
- Balog, K. 1999. Conceivability, possibility, and the mind-body problem. *Philosophical Review* 108:497-528.
- Block, N. & Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review* 108:1-46.
- Byrne, A. 1999. Cosmic hermeneutics. *Philosophical Perspectives* 13.
- Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D.J. 1999. Materialism and the metaphysics of modality. *Philosophy and Phenomenological Research* 59:473-96.
- Chalmers, D.J. (forthcoming a). The content and epistemology of phenomenal belief. In (Q. Smith & A. Jokic , eds) *Aspects of Consciousness*. Oxford University Press.
[www.u.arizona.edu/~chalmers/papers/belief.html]
- Chalmers, D.J. (forthcoming b). Does conceivability entail possibility? In (T. Gendler & J. Hawthorne, eds) *Imagination, Conceivability, and Possibility*. Oxford University Press.
[www.u.arizona.edu/~chalmers/papers/conceivability.html]
- Chalmers, D.J. (forthcoming c). The nature of epistemic space.
[www.u.arizona.edu/~chalmers/papers/espace.html]
- Chalmers, D.J. (forthcoming d). The foundations of two-dimensional semantics. In (M. Garcia-Carpintero & J. Macia, eds) *Two-Dimensionalism*. European Review of Philosophy.
[www.u.arizona.edu/~chalmers/papers/foundations.html]
- Hill, C.S. & McLaughlin, B.P. 1999. There are fewer things in reality than are dreamt of in Chalmers' philosophy. *Philosophy and Phenomenological Research*.

- Horgan, T. 1984. Supervenience and cosmic hermeneutics. *Southern Journal of Philosophy Supplement* 22:19-38.
- Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-36.
- Jackson, F. 1994. Armchair metaphysics. In (J. O'Leary-Hawthorne & M. Michael, eds) *Philosophy in Mind*. Kluwer.
- Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.
- Jackson, F., Pettit, P. & Smith, M. 2000. Ethical particularism and patterns. In (B. Hooker & M. Little, eds) *Moral Particularism*. Oxford University Press.
- Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 4:435-50.
- Levine, J. 1993. On leaving out what it's like. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.
- Levine, J. 1998. Conceivability and the metaphysics of mind. *Nous* 32:449-480.
- Lewis, D. 1994. Reduction of mind. In (S. Guttenplan, ed.) *A Companion to the Philosophy of Mind*. Oxford: Blackwell.
- Loar, B. 1997. Phenomenal states (second version). In (N. Block, O. Flanagan, and G. Güzeldere, eds) *The Nature of Consciousness: Philosophical Debates*. MIT Press.
- Papineau, D. 1993. Physicalism, consciousness, and the antipathetic fallacy. *Australasian Journal of Philosophy* 71:169-83.
- Salmon, N. 1986. *Frege's Puzzle*. MIT Press.
- Searle, J.R. 1991. *The Rediscovery of the Mind*. MIT Press.
- Yablo, S. 2000. Textbook Kripkeanism and the open texture of concepts. *Pacific Philosophical Quarterly* 81:98-122.
- Yablo, S. (forthcoming). Coulda, woulda, shoulda. In (T. Gendler & J. Hawthorne, eds) *Imagination, Conceivability, and Possibility*. Oxford University Press. [www.mit.edu/~yablo/coulda.pdf]

Materialism and the Metaphysics of Modality

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

This appeared in *Philosophy and Phenomenological Research* 59:473-93, as a response to four papers in a symposium on my book *The Conscious Mind*. Most of it should be comprehensible without having read the papers in question. This paper is for an audience of philosophers and so is relatively technical. It will probably also help to have read some of the book. (There is a corresponding [precis](#) of the book, written for the symposium.) The papers I'm responding to are:

- Chris Hill & Brian McLaughlin, There are fewer things in reality than are dreamt of in Chalmers' philosophy
- Brian Loar, David Chalmers' *The Conscious Mind*
- Sydney Shoemaker, On David Chalmers' *The Conscious Mind*
- Stephen Yablo, [Concepts and consciousness](#)

Contents:

- 1 [Introduction](#)
- 2 [Logical supervenience](#) (Shoemaker)
- 3 [Type-B materialism](#)
 - 3.1 [The form of the argument](#) (Hill/McLaughlin, Loar)
 - 3.2 [Are there logically possible zombie worlds?](#) (Yablo, Loar)
 - 3.3 [Are there strong necessities? I. Examples](#) (Yablo)
 - 3.4 [Are there strong necessities? II. Explanations](#) (Hill/McLaughlin, Loar)
 - 3.5 [Modal rationalism](#)
 - 3.6 [Type-B materialism: Loose ends](#) (Hill/McLaughlin, Loar)
- 4 [Epiphenomenalism and the paradox of phenomenal judgment](#) (Hill/McLaughlin, Shoemaker)
- 5 [Dancing qualia](#) (Shoemaker)

1 Introduction

The argument against materialism in *The Conscious Mind* has two parts. The first part, in Chapter 3 of the book, argues that facts about consciousness are not necessitated *a priori* by physical facts. The second part, in Chapter 4 of the book, argues that facts about consciousness are not necessitated *a posteriori* by physical facts. Either part of the argument might be resisted. Corresponding to these paths of resistance,

there are two very different brands of materialism, which I call type-A and type-B materialism.

Type-A materialists hold that phenomenal facts (insofar as there are such facts) are necessitated *a priori* by physical facts. Such a materialist denies that physically identical zombie worlds or inverted-qualia worlds are coherently conceivable, denies that Mary (of the black-and-white room) gains any factual knowledge on seeing red for the first time, and typically embraces a functional (or eliminative) analysis of consciousness.

Type-B materialists accept that phenomenal facts are not necessitated *a priori* by physical facts, but hold that they are necessitated *a posteriori* by physical facts. Such a materialist accepts that zombie worlds or inverted-qualia worlds (often both) are coherently conceivable but denies that such worlds are metaphysically possible, holds that the factual knowledge that Mary gains is knowledge of an old fact in a new way, and typically embraces an *a posteriori* identification of consciousness with a physical or functional property.

All of the commentators in this symposium are type-B materialists, and three of the commentaries focus on my argument against an *a posteriori* necessary entailment. This makes for a focused and productive discussion. I would like to thank all the commentators for their detailed, sophisticated, and generous commentaries, which in all cases significantly advance the discussion of important issues.

I will organize my replies to parallel the discussion in the book. In order, I will discuss issues relevant to the argument against *a priori* entailment (Chapter 3), then those relevant to the argument against *a posteriori* entailment (Chapter 4, backed by Chapter 2), then those relevant to epiphenomenalism and the paradox of phenomenal judgment (Chapter 5), and finally those relevant to the fading and dancing qualia arguments for nonreductive functionalism (Chapter 7). The first, third, and fourth of these parts mostly revolve around Shoemaker's commentary. The second part, and by far the longest, mostly revolves around the commentaries by Hill & McLaughlin, Loar, and Yablo.

2 Logical supervenience

All of the commentators are type-B materialists, holding that the phenomenal facts are not entailed *a priori* by the physical facts, or do not logically supervene on those facts, as I put in in the book. Unlike the others, Shoemaker endorses one element of the type-A materialist position, in that he denies that zombies are conceivable. He allows that inverted spectra are conceivable, which is enough to defeat logical supervenience in general. In effect, he holds that *some* phenomenal facts (facts concerning the existence of consciousness in a system, and facts about qualitative similarity) logically supervene on physical facts, but others (facts about the specific nature of experiences) do not. Facts in the second class are entailed only *a posteriori* by physical facts.

Shoemaker does not argue directly for his position here. But he says that its coherence threatens my arguments in two places. First, he says it shows how there could be a necessary entailment from physical to phenomenal even without an *a priori* entailment. Second, he says it shows that I am not "entitled" to

the claim that zombies are conceivable.

Taking the second point first: Shoemaker is correct to say that the conceivability of zombies is not implied by the failure of *a priori* entailment. But he is incorrect in saying that I suppose that it is so implied, and in suggesting that my entitlement to the former rests on my entitlement to the latter. The direction of support is the other way around: the conceivability of zombies is used as one argument for the failure of logical supervenience. (There are other arguments for the failure of logical supervenience, and I note (p. 101) that one might endorse this failure without endorsing the conceivability of zombies.) The argument for the conceivability of zombies rests on quite distinct considerations.

Shoemaker may be misled by the sentence he quotes from Chapter 1 (p. 23): "If there were a functional analysis of the notion of experience or of phenomenal quality, then the analysis in question would yield functional analyses of specific phenomenal properties." I take the blame for misleading him here, as the sentence is ambiguous. If "the analysis in question" is taken to refer to the analysis in the first part of the sentence, this claim might be questionable. But the expression was in fact intended to refer to a quite different analysis under discussion earlier in that paragraph. The "analysis in question" is the analysis of specific phenomenal properties as "the sort of conscious experience that tends to accompany [functional property] P". If there were a functional analysis of the concept of conscious experience, one could plug it into the quoted analysis to yield a functional analysis of the specific phenomenal properties in question, as I asserted. I apologize for the misleading ambiguity.

Shoemaker also suggests that his position shows how a psychophysical entailment can be necessary without being *a priori*. I don't think this is quite right, however. Even the conceivability of inverted worlds is enough for my arguments against a necessary connection (in Chapter 4) to go through. If we combine the conceivability of such worlds with the two-dimensional analysis of *a posteriori* necessity, it turns out that there can be no *a posteriori* necessary connection from physical facts to specific phenomenal facts (essentially because specific phenomenal concepts will have primary intensions that do not supervene on the physical), and Shoemaker's position will be ruled out. Of course Shoemaker may hold that there are *a posteriori* necessities that do not fit the two-dimensional model; but that point needs independent argument, and it is that argument that will do all the work. Merely exhibiting the position does no work, as my argument in Chapter 4 is itself an argument against the position.

Of course I disagree with Shoemaker's position on independent grounds. In particular I don't think that qualitative similarity and difference are functionally definable (partly because of the conceivability of zombies, and partly because I think that qualitative similarity depends only on intrinsic properties of experiences). Shoemaker says his claim should be "no more controversial" than my claim that judgment, memory, and other psychological states are functionally definable; but the claim about "judgment" and "psychological states" involve stipulative uses of those terms and are therefore trivial, while the claim about memory is not made in any definitive way and plays no role in my arguments. Even if one accepts his claim, I don't think it implies that specific qualia are fixed by physical facts, as he suggests. Shoemaker's discussion of "realizing" functional definitions seems to suggest that specific qualia should be rigidly defined as the occupiers of the relevant functional role; but that definition would be incompatible with the failure of *a priori* entailment. And without such a definition, there is no reason to

think that fixing qualitative similarity fixes qualia. In any case, the failure of *a priori* entailment (which Shoemaker concedes) is all that my central arguments require.

This touches on a minor misrepresentation of my argument in Chapter 3 that occurs repeatedly. Loar says my argument against *a priori* entailment is "grounded" in the conceivability of zombies, and Yablo says "everything turns on" the claim that there are zombie worlds. In fact the zombie argument is just one of five arguments against *a priori* entailment, and is not obviously the most important. First, the inverted spectrum and Mary arguments make claims that do not imply the zombie claim, so someone (like Shoemaker) who rejects zombies may still accept the first two. Second, it may be more correct to say the zombie argument is grounded in the argument from absence of structural/functional analysis than vice versa. Finally, at least from a rhetorical standpoint, the Mary argument works better than the zombie argument against a type-A materialist (a conceivability claim is easy to deny, a new-knowledge claim much harder). So I would not like to give the impression that everything depends on zombies here.

3 Type-B materialism

Type-B materialism has been perhaps the most popular position on consciousness in recent years. It simultaneously promises to take consciousness seriously (avoiding the deflationary excesses of type-A materialists) and to save materialism (avoiding the ontological excesses of the property dualist). For these purposes, Kripke's introduction of *a posteriori* necessities has seemed a godsend. In recent years, almost every type-B materialist (Loar is a notable exception) has appealed to Kripkean examples in support of their position. On such a view, the relationship between consciousness and physical processes is supposed to be like that between water and H₂O, or between Hesperus and Phosphorus.

In *The Conscious Mind* I argue that this does not work. These (and other) standard examples of *a posteriori* necessity give no support to a type-B materialist. Kripke's *a posteriori* necessities can all be seen to fall under a two-dimensional framework, which has just one space of worlds (the conceivable worlds) and two intensions (primary and secondary) for terms and statements over those worlds. On this framework, the distinction between conceivability and possibility, and between epistemic and metaphysical necessity, falls at the level of statements, not worlds. In particular, *a posteriori* necessities are statements with a contingent primary intension and a necessary secondary intension. It is relatively easy to see that necessities of this sort cannot save the materialist. So if *a posteriori* necessity is to save the materialist, it must be a radically new form of *a posteriori* necessity, a form not explicable by the two-dimensional framework, and not supportable by appeal to the Kripkean examples.

Little of this is disputed by the current commentators. Instead most of them bite the bullet and argue that psychophysical necessities are different in kind from the Kripkean examples, and not explicable by the two-dimensional framework. Let us call these *a posteriori* necessities not explicable by the 2-D framework *strong necessities*. I argue in the book that there is no reason to believe in strong necessities, and that they are highly problematic. In response, Hill & McLaughlin and Loar offer what they take to be accounts of the roots of these strong necessities; and Yablo gives reasons why he thinks strong necessities must exist. Before discussing these crucial counterarguments, I will first clarify some issues concerning

my argument.

3.1 The form of the argument

Hill & McLaughlin and Loar focus on the four-step argument at the beginning of Chapter 4. This formulation does not really engage the type-B materialist, however, as it was intended to be accessible to readers ignorant of the purported distinction between conceptual and metaphysical necessity. The material most relevant to a type-B materialist can instead be found in section 2 of chapter 4, on objections from *a posteriori* necessity, and also in section 4 of chapter 2, which outlines the 2-D framework.

As far as the four-step argument is concerned, Loar is correct that the jump from (2) to (3) is the part that a type-B materialist should dispute. Hill & McLaughlin dispute (2) instead, but this rests on a misunderstanding of my use of "logical possibility". On my usage (as Loar and Yablo note), "logically possible" is near enough to interchangeable with "conceptually possible" (*not* "metaphysically possible"), and is tied by stipulation to conceivability.[*] Without noting this usage, many passages and arguments will seem nonsensical.

*[[Throughout this discussion, "conceivability" should be understood as ideal conceivability, or conceivability on ideal rational reflection.]]

Here is a crucial passage from p. 68 of the book:

Most importantly, none of the cases we have seen give reason to believe that any conceivable *worlds* are impossible. Any worries about the gap between conceivability and possibility apply at the level of *statements*, not worlds: either we use a statement to misdescribe a conceived world (as in the Kripkean case, and the second Goldbach case mentioned), or we claim that a statement is conceivable without conceiving of a world at all (as in the first Goldbach case). So there seems to be no reason to deny that conceivability of a world implies possibility. I will henceforth take this for granted as a claim about logical possibility; any variety of possibility for which conceivability does not imply possibility will then be a narrower class. Someone might hold that there is a narrower variety of "metaphysically possible worlds"; but any reason for believing in such a class would have to be quite independent of the standard reasons I have considered here.

Perhaps it was not clear enough that "taking for granted" here is stipulative. I note, though, that the notion of logical possibility is introduced on p. 35 with a tie to conceivability, and the quoted passage (and much of the rest of the book) does not really make sense otherwise. I also don't think my usage is unusual; it's typical in philosophy for the adjective "logical" to be tied to matters of conceptual truth and rational inference. I don't use "conceptual possibility" as the term doesn't seem quite right to me: it downplays the role of modal intuition (as opposed to mere semantic intuition); and while mathematical truths are certainly logically necessary on my usage, it is widely held that many of them are not conceptually necessary. But if a reader wants to substitute "conceptual" for "logical" on every occurrence, she will not

be led too badly astray.

Of course, I hold that conceptual possibility = logical possibility = metaphysical possibility (at the level of worlds). But when we are discussing a potential *distinction* between conceptual and metaphysical possibility, "logical possibility" always goes with the former and not with the latter. Hill & McLaughlin read it as going with the latter, hence the confusion.[*]

*[[Hill's commentary on my book in *Analysis* (Hill 1998) rests largely on this misunderstanding. Hill says that my claim that we have *a priori* access to the space of logically possible worlds "begs the question"; but when "logical possibility" is understood appropriately, the claim becomes relatively uncontroversial.]]

(Two further clarifications: Hill & McLaughlin locate my "argument" for premise (2) in an incidental passage on p. 131, but in fact it is given in detail in Chapter 3. And they say that I hardly discuss the view that phenomenal states are *a posteriori* identical with neurobiological states, but in fact it is a special case of the type-B materialist view that I discuss at considerable length.)

With this matter clarified, the step that a type-B materialist must question is certainly the step from the failure of logical supervenience to the failure of metaphysical supervenience (and hence the falsity of materialism). My argument for this step has three parts. The first part (section 4 of chapter 2) outlines the two-dimensional account of *a posteriori* necessity and argues that it handles all standard *a posteriori* necessities without a gap between conceivable and possible worlds. The second part (section 2 of chapter 4) applies the 2-D account to argue that *a posteriori* necessities of the usual variety cannot save materialism. The third part (later in the same section) argues that there is no reason to believe in "strong necessities" which escape the 2-D account.[*]

*[[Yablo calls the view that all *a posteriori* necessities fall under the two-dimensional account "textbook Kripkeanism". I think this may be the wrong term, partly because it is not clear that Kripke accepts the view himself, and partly because many textbook uses of *a posteriori* necessity are incompatible with it. If one needs a name for the position, I might suggest "two-dimensionalism" or perhaps "modal rationalism".]]

The first two parts of the argument are straightforward, and none of the commentators dispute them. But a point of clarification. Hill & McLaughlin and Loar say that (in the second part) I rely on the Kripke-style observation that when it comes to phenomenal concepts, primary intension and secondary intension are the same. In fact (as Yablo notes), while I accept this observation, it is inessential to the argument. Even if primary and secondary intensions of phenomenal concepts differ, the argument against materialism goes through, although things have to be put slightly differently.

It is easy to see how, *if* we accept the 2-D account of *a posteriori* necessity, any apparent "holes" in my argument are filled in.[*] Loar's question about the step from distinct primary intensions to distinct properties is answered straightforwardly, for example: such intensions will provide different functions from worlds to extensions (remember, there is just one space of worlds), and therefore distinct properties (wrinkles about centering aside). Hill & McLaughlin question the same move, but get into unnecessary difficulties. The passage that they quote and criticize from p. 373 in fact ends as follows:

*[[There is one loophole that remains. One can appeal to a distinction between primary and secondary intensions for fundamental *physical* concepts, postulating phenomenal or protophenomenal grounds for basic physical dispositions (see pp. 134-6). It is not entirely clear that the resulting "panprotopsychism" counts as materialism, but in any case it will be much more to my tastes than to those of most materialists.]]

if the subject cannot know that *R* is *P a priori*, then reference to *R* and *P* is fixed in different ways and the reference-fixing intensions can come apart in certain conceivable situations. Unless we invoke the additional machinery of strong metaphysical necessity, the difference in primary intensions will correspond to a difference in reference-fixing properties.

Nothing Hill & McLaughlin say provides any reason to deny this. They change "reference-fixing intensions" in their quotation to "[properties]", but the change is not innocent. The claim in the first sentence involves primary intensions across conceivable worlds, and is relatively uncontroversial; even Loar accepts it. And the claim in the second sentence is that this difference in intensions yields a difference in properties *unless* one accepts strong necessities (i.e., unless one holds that there are *a posteriori* necessities that escape the 2-D framework). Again, there is little reason to deny this.[*]

*[[Note that my claims in this passage require none of Hill & McLaughlin's premises (a), (b), and (c). Hill & McLaughlin think I need (a) because they overlook the significance of their substitution; they think I need (b) and (c) because they overlook the significance of Loar's term "introduce" in the quoted passage. As for my overall argument, I do require (a), but (b) and (c) are inessential.]]

The residual question, then, is whether there are strong psychophysical necessities. Standard *a posteriori* necessities have a necessary secondary intension but a contingent primary intension. A strong necessity, by contrast, is an *a posteriori* necessity with a necessary primary intension. (This should be taken as the defining property of a strong necessity.) The primary intension of a strong necessity (unlike that of a standard *a posteriori* necessities) will be false in no metaphysically possible world. Of course the primary intension will be false in some *conceivable* world, because the statement is *a posteriori*. Hence the crucial feature of strong necessities, which is that they require a distinction between conceivability and possibility at the level of worlds.

In the book (p. 136), I call these "strong metaphysical necessities", and characterize them by the view that there are fewer metaphysically possible worlds than logically possible worlds. Not all of the commentators recognize their view in this description. Hill and McLaughlin's resistance is explained by their misreading of my use of "logically possible". (They also reject my characterization of the view as requiring that such necessities stem from factors independent of the semantics of the terms involved; I should have said "from the two-dimensionality of the semantics of the terms involved".) Yablo resists by denying that zombie scenarios involve "worlds" at all. Before addressing the question of whether there are strong necessities, I will address this point of Yablo's.

3.2 Are there logically possible zombie worlds?

In an interesting line of resistance, Yablo accepts that it is logically (or conceptually) possible that there are zombies, but denies that there are logically possible zombie worlds. One way to make this denial would be to stipulate that "world" means "metaphysically possible world". This would make the denial a mere terminological variant of the usual core issue: instead of discussing a gap between conceptually possible and metaphysically possible worlds, we would have to discuss a gap between conceptually possible "scenarios" and metaphysically possible worlds, where scenarios are world-like except that they are metaphysically impossible. But I think Yablo is making a deeper point, questioning whether there are any world-like objects that correspond to "zombie worlds" at all.

I think the answer to this question is clearly yes: if zombies are ideally conceivable, we need logically possible zombie worlds.[*]

*[[The same goes for inverted worlds. Here and in what follows, I use zombie worlds to stand in for any world physically the same as and phenomenally different from our own. If we allow that the primary and secondary intensions of phenomenal concepts may differ, this should be taken to include worlds in which the primary intensions of our phenomenal concepts yield different results.]]

To consider the issue, pretend for a moment that zombies are conceivable but not metaphysically possible. Even better, take another "strong necessity" view, on which the only possible worlds are those whose laws of nature are our laws. (This view differs from a weaker view on which apparently counternomic worlds should be redescribed as not containing "mass" and such at all. The latter view also endorses the metaphysical necessity of laws, but is compatible with the 2-D framework.) On such a view, counternomic scenarios, although conceivable, are reflected in no metaphysically possible worlds at all. But we still need *logically* possible worlds with different laws, for all sorts of reasons.

Think of the reasons why "possible worlds" talk is introduced into philosophy in the first place. Possible worlds are introduced to deal with counterfactual thought, the semantics of counterfactual language, rational inference, and the contents of belief, among other reasons. A scientist can think counterfactually (and rationally) about scenarios with different laws, and can make true utterances about these scenarios. If we are to use possible-worlds talk to characterize the contents of her beliefs in discovering laws, we will need to appeal to counternomic worlds. Without counternomic worlds, we will not be able to use worlds to make sense of her inference processes. And so on. Ruling out counternomic worlds will make possible worlds useless for many or most standard purposes. Even if someone insists that such worlds are not metaphysically possible, we *need* logically possible counternomic worlds.

Something similar applies to zombie worlds. Even on a type-B materialist view, we can think counterfactually (and rationally) about the possibility of a different distribution of phenomenal properties with the same physical properties. We need worlds corresponding to these possibilities to make sense of counterfactual thought, of the semantics of counterfactual utterances, of rational inferences involving consciousness, of the contents of rational beliefs about consciousness, and so on. We can write coherent science fiction about zombies, and speak coherently about the truth in such fictions. Talk of logically possible zombie worlds is justified in the usual way by their role in these uses.

I suppose one might resist by holding an extreme modal-realist view (even stronger than David Lewis's) on which possible worlds are simply "out there" and on which their connection to the rational uses in question is neither here nor there. But even on such a view, we would end up having to postulate worldlike objects ("ersatz" worlds, at the very least) for the rational purposes. And on the dominant view on which all counterfactual worlds are regarded as ersatz abstract objects, there seems no ground for resistance. It is easy to construct an ersatz object that behaves in just the way that a zombie world should. The obvious strategy is to use maximal consistent worldbooks, where "consistent" is understood in the *a priori* sense. One has to tread carefully in handling two-dimensional phenomena and centering, but the matter is straightforward.

Another line of resistance might liken zombie worlds to supposed "impossible worlds" with all their problems, but the analogy doesn't work. Impossible worlds simply don't behave as worlds should: statements are both true and false there, for example. In zombie worlds, statements can be semantically evaluated with ease, and they never come out both true and false. Zombie worlds can even be considered both as actual and as counterfactual, as a world should be, and statements will be well-behaved on both methods of evaluation.

Loar suggests a concern: if zombie worlds are logically possible but not metaphysically possible, such worlds won't satisfy the necessity of identity. I think this is a better argument against strong necessities than against zombie worlds; what causes problems is not the zombie worlds themselves but the postulation of strong necessities and the associated problematic identities. But if one does hold the strong necessity view, I think the right thing to say is that the necessity of identity applies only across metaphysically possible worlds, not logically possible worlds. At worst, if someone insists both on strong necessities and that worlds must satisfy the necessity of identity, we'll just have to call the zombie situation a "scenario" or some such instead. Nothing crucial changes.

So, it seems that there are overwhelming grounds for admitting something like zombie worlds into our ontology, at least given that zombies are ideally conceivable and that we have worlds in our ontology at all. More generally, there are overwhelming grounds for admitting a logically possible world for every ideally conceivable scenario: we need to do so in order to make sense of most of the things that possible worlds are supposed to make sense of. The only question is whether we *also* need to stipulate a distinct modality of metaphysical possibility, holding that only some of these worlds are metaphysically possible. I think not, but more on this shortly.

I have in effect just argued for Yablo's CP: we need a logically possible world for every ideally conceivable scenario if we are to make sense of the various rational notions in question. Yablo's (1)-(7) argument for CP is not my argument: on my reading, his (2) is weaker than his (1) and his (5), for example. Yablo is right, though, that I need to be able to move from *de dicto* to *de re* possibility where logically possible worlds are concerned, and that I need to exclude opacity phenomena. The conception of logically possible worlds I am using does just that. With a world for every ideally conceivable scenario, it follows that whenever it is conceivable that *P*, there will be a logically possible world in which *P* (according to primary intension). And whenever two conceived-of scenarios are not identifiable even

under ideal rational reflection, they will correspond to distinct logically possible worlds.

Yablo asks how we conceive of worlds in order to evaluate them under primary intensions. I take Yablo's first option: we conceive of them under descriptions. I certainly don't endorse the "immaculate conception" option. Perhaps a couple of my remarks suggest the "what I would say if plopped down in W" option; this may be a useful heuristic but it is not my considered view, partly for the reasons Yablo mentions. Instead, I take a (centered) world - always under a description, of course - and consider it as an epistemic possibility, asking "what is the correct thing to say if this world is actual".

I avoid opacity phenomena not by conceiving under canonical descriptions (as Yablo suggests in his discussion of the first option), but simply because rationally inequivalent descriptions will pick out inequivalent worlds, by definition of logically possible worlds. Of course a world can be described in many different ways, but such descriptions will always yield the same results when primary intensions are evaluated, as the equivalence between the descriptions will be revealed on ideal rational reflection.

(This applies at least when the descriptions under which we conceive of worlds are intended as "primary" descriptions. We can also conceive of worlds under "secondary" descriptions, if we like, but then all bets are off. (This is more or less the distinction between 1-conceivability and 2-conceivability outlined in the book, or Yablo's distinction between conceiving of an E-verifying world and an E-satisfying world.). When worlds are conceived of in the secondary way, opacity phenomena can arise: we won't in general know whether two descriptions describe the same world (unless we have relevant empirical knowledge of the actual world, or unless the concepts' primary and secondary intensions coincide), and we can't always know whether we have conceived of a world. When worlds are conceived of in the primary way, however, no such problems arise.)

3.3 Are there strong necessities? I. Examples

The discussion of strong necessities in my book is relatively brief, as few people in the literature had embraced them; the great majority of type-B materialists appealed to the Kripkean examples instead. (Loar is an exception.) But I argue against strong necessities (pp. 136-38), saying there is no reason to believe in them. In particular, I say that (a) they cannot be supported by analogy with other *a posteriori* necessities; (b) they involve a far more radical sort of *a posteriori* necessity than Kripke's, requiring a distinction between logical and metaphysical possibility at the level of worlds; (c) they lead to an *ad hoc* proliferation of modalities, (d) they raise deep questions of coherence, (e) strong necessities will be brute and inexplicable, and (f) the only motivation to postulate such necessities is the desire to save materialism. I stand by these points, but there is more to say about all of them, and each of the commentaries raises interesting questions.

The first part of the argument is the absence of analogy: there are no other counterexamples to the two-dimensional account of *a posteriori* necessity. I think it is relatively uncontroversial that *most a posteriori* necessities fall under the 2-D model, and certainly that the most commonly accepted examples do. Loar accepts explicitly that there are no other counterexamples to the 2-D model: he thinks that psychophysical

necessities are *sui generis*. Hill & McLaughlin appear to accept the same thing. Shoemaker touches on one source of counterexamples, though, and Yablo discusses a potential counterexample in detail.

Shoemaker mentions his view that the laws of nature are metaphysically necessary. As we have seen, this view can be interpreted as requiring that laws are strong necessities. These purported necessities are at least as tendentious as psychophysical necessities (and far less widely accepted), however, and I would use the 2-D framework to argue equally against them.

The same goes for the claim that God exists necessarily, which would also provide a strong necessity if true. Yablo adapts the God example to provide an ingenious argument for strong necessities, however. In effect, he notes that one of the following statements must be necessary: "a necessarily existing god exists", and "a necessarily existing god does not exist". But each of these statements is conceivably false (he suggests), and therefore *a posteriori*; and the necessary *a posteriori* status of either statement is not plausibly explicable by the 2-D framework. So one of these statements is a strong necessity.

I deny this, because I deny that a necessarily existing god is conceivable. Such a god may be conceivable in the sense of "not obviously inconceivable", but in no stronger sense. I certainly can form no clear and distinct conception of such a god (like many, I was suspicious of the idea the moment I heard about it as a student), and continued rational reflection reveals all sorts of problems with the idea. One thing that rational reflection reveals is the conceivability of a world *without* a god. If such an intuition is accepted, it causes grave difficulty for the coherence of the idea of a necessarily existing god. (Ideal reflection allows us to use one modal intuition in evaluating others, of course.)

The problematic issues here arise because of the double modality: we are conceiving not just of worlds "in themselves", but also of what is possible or necessary within those worlds. Conceiving of a god (an omnipotent, omniscient, and benevolent being, say) is arguably not too hard; but to conceive in addition that the being exists necessarily, we have to conceive that the space of possible worlds is such that this god exists in each of them, *despite* the conceivability of a godless world. That is, we have to conceive that the conceivability/possibility thesis (CPT: all conceivable worlds are possible worlds) is itself false. This is what does all the work in the example: if it is conceivable that CPT is false, then (by CPT!) it is possible that CPT is false, and CPT is therefore false (as it is surely necessarily true, if true).

A natural response to this sort of argument is to restrict the conceivability/possibility thesis to claims about the distribution of (nonmodal) properties within worlds, leaving double modals outside its scope. I think this response would be defensible, and not entirely *ad hoc* (CPT might still apply to worlds in themselves, although not to "cosmoses" of possible worlds; all that changes is that one can't make Yablo's "*de dicto* to *de re*" step for double modals). But I prefer to hold onto the stronger thesis, by denying that it is conceivable that CPT is false. I hold that CPT is *a priori*, although highly nontrivial, like many theses in philosophy. In fact I will sketch an *a priori* argument for CPT later in this paper. If this is correct, then CPT is not conceivably false on ideal rational reflection, and it is not ideally conceivable that a necessarily existing god exists.

One might resist the conclusion by stripping down one's conception of a god so that a god's nonexistence

becomes inconceivable. If we conceive of a god (nonrigidly) as "all there is", for example, then a god exists necessarily. (My discussion above assumes that a god is required to have contingent-seeming properties such as omnipotence.) The case of numbers and sets, which Yablo raises, is something like this. I think it is *a priori* that numbers exist, so inconceivable that they do not exist. Of course this depends on one's conception of numbers. There are alternative conceptions on which it is *a posteriori* that numbers exist, or on which it is *a priori* that numbers do not exist. What matters is that there is no tenable conception of numbers on which (a) there is a conceivable world in which numbers exist, (b) there is a conceivable world in which numbers do not exist, and (c) if numbers exist, they exist necessarily. I think there is no reason to deny this.[*]

*[[Another class of candidates for strong necessities might include mathematical statements that are necessarily true (even by primary intension) but not knowable *a priori*. It is far from clear that such statements exist (idealizing away from our cognitive limitations), but even if they exist, I do not count them as strong necessities. Unlike strong necessities, such statements open no gap between the conceivability and possibility of worlds, as opposed to statements; there is no conceivable world (considered as actual) in which these statements are false. I think such statements, if they exist, are best regarded as neither *a priori* nor *a posteriori*, but as a third category, "inscrutable". (The epistemic theory of vagueness, if true, would supply further inscrutable truths.) The existence of inscrutable truths would not help the type-B materialist, who accepts that zombie scenarios are clearly conceivable, so I set them aside here. I discuss this topic in more depth in forthcoming work.]]

3.4 Are there strong necessities? II. Explanations

If I am right so far, there are no good reasons to believe in strong necessities outside the psychophysical case. But Hill & McLaughlin and Loar both provide interesting potential *explanations* of why there should be strong necessities in the psychophysical case. If this worked, it would be as good as a resolution of the mind-body problem as any. Unfortunately, I don't think the proffered explanations work. Below I explain why.

Hill & McLaughlin offer a psychological explanation of why we can conceive of zombies, in terms of the independent cognitive processes by which we conceive of physical processes and by which we conceive of experiences. There are some questions about the details of this explanation; and I should note that it is not obviously a purely physical/functional explanation, as it helps itself to the unreduced notion of a sensory concept. But let me grant that some explanation along these lines can be given. The question is whether it provides any explanation of strong necessities. On the face of it, it seems not. After all, one can also give a psychological explanation of why we can conceive of red squares, in terms of the distinct cognitive processes and epistemic constraints involved in conceiving of color and shape. One can give a psychological explanation of why we can conceive of five-horned animals, or of silicon-based life. But no-one would infer that there are strong necessities denying the metaphysical possibility of red squares or five-horned animals or silicon-based life.

The form of Hill & McLaughlin's argument is far from unique to the psychophysical domain. The precise details (concerning sensory concepts and recognitional concepts) may be unique, but these details are inessential. What does the work in their argument is the distinctness of the "faculties" or the "epistemic

constraints" that these concepts involve. One could make an argument with the same form about shape concepts and color concepts, an argument that should then be just as compelling. And one could make a very similar argument concerning any two pairs of concepts with different cognitive roles, thus "explaining away" any conceivability intuition at all. (There will always be a cognitive explanation of a modal intuition!) If this account tells us that Cartesian modal intuitions are unreliable, the same will go for all modal intuitions.

(Worse: a similar account might tell us that all of our mathematical beliefs are unreliable. There will presumably be cognitive explanations of these beliefs, too.)

The trouble is that an explanation of a strong necessity has to do two things: it has to show us why a state-of-affairs should be conceivable while at the same time being impossible. To put matters differently, it should explain why conceivability is an *unreliable* guide to possibility of such states-of-affairs (Hill & McLaughlin put it this way themselves). But Hill & McLaughlin have discharged only half of the burden, and the easy half: they have explained why zombies might be conceivable. But they have done nothing to explain why we should take this conceivability to be unreliable.

Hill (1997) likens his strategy to Kripke's "explaining away" of modal intuitions about heat and molecular motion, but I think it is crucially different. Kripke's strategy explains why necessity and apparent contingency are *co-present*: if we mistakenly describe a conceived situation by considering it as actual rather than as counterfactual, we should *expect* that certain necessary statements will falsely appear to be contingent. By contrast, Hill's strategy merely explains the appearance of contingency; why this appearance is co-present with necessity is not explained at all. At most, Hill tacks the hypothesis of necessity onto the end of his explanation of contingency, so it remains an unexplained explainer.

The closest Hill & McLaughlin come to addressing this burden is saying: "Given these differences between sensory concepts and physical concepts, a sensory state and its nomologically correlated brain state would seem contingently connected *even if they were necessarily one*" (my italics). But this may well be a deeply "*per impossibile*" counterfactual ("given mathematical concepts, $1+1$ would seem to be 2 even if it were 3"). What we need is an explanation of how the two states *could* be necessarily one. Or ascending to the level of concepts, we need an explanation of how two such distinct concepts *could* pick out the same property, and indeed have the same property as a reference-fixer.

That is just what Loar attempts to provide. Some parts of Loar's discussion appeal only to the distinctness of cognitive role of the two sorts of concepts; as an "explanation", this has the same problems as Hill & McLaughlin's. But other parts attempt something more. In effect, they attempt to explain strong necessities by explaining how it is that physical and phenomenal concepts could pick out the same property and have the same property as their reference-fixer, despite their cognitive distinctness. The details of Loar's account here are somewhat vague, but they are spelled out in much more detail in Loar (1997), which I rely on below.

Loar appeals to two facts about phenomenal concepts: they are (a) recognitional concepts[*] that (b) express the same property that they refer to. (In Loar's terminology, the property a concept "expresses" is

its reference-fixer, or the property corresponding to its primary intension across possible worlds.) The significance of (a) is that recognitional concepts and physical-theoretical concepts are generally cognitively distinct even when they refer to the same property. But this is not enough to save materialism: other recognitional concepts conceive of their referents under contingent modes of presentation (i.e., they have distinct primary and secondary intensions across possible worlds), so they *express* a property quite distinct from that expressed by the corresponding theoretical property. So Loar adds (b), noting that in the phenomenal case, uniquely, we have recognitional concepts that express the property they refer to. He appeals here to Kripke's observation about pain, in effect noting that phenomenal concepts have the same primary and secondary intension, and do not conceive of their referents under contingent modes of presentation. Granting all this, we deduce that the two concepts are cognitively distinct and coreferential (because of (a)), and that each concept expresses the property it refers to[*] (because of (b)), from which it follows that the two concepts have the same reference-fixing property.[*] So strong psychophysical necessities are explained.

*[[I am not sure that phenomenal concepts should be thought of as recognitional concepts, but I will go along with Loar on this point here. In his paper in this symposium, Loar sometimes says "experiential" rather than "recognitional", but he means something similar. "Experiential concept" has a different meaning for me (it is more or less synonymous with "phenomenal concept"), so I will use "recognitional" here.]]

*[[This also requires the assumption (implicit in Loar's discussion) that a physical-theoretical concept expresses the same property that it refers to. I think this is arguable, but I will go along with it to keep things simple. If this assumption is retracted, related versions of Loar's point and my reply can still be given.]]

The problem with this is straightforward. The introduction of point (b) undercuts the value of point (a). The significance of point (a) for Loar lies not just in the cognitive distinctness but also the coreference of recognitional and theoretical concepts. But this coreference is *explained* by the two-dimensional nature of such recognitional concepts: they typically conceive of their referent as "the cause of such-and-such experience", or under some similar contingent mode of presentation. If we remove this feature of recognitional concepts (as we do in accepting (b)), we no longer have any reason to believe that recognitional concepts and distinct theoretical concepts should corefer.

As things stand, in accepting (a) and (b) we are left with the observation that phenomenal concepts and physical concepts (i) are cognitively distinct, and (ii) both express the property that they refer to. It's clear that nothing here begins to justify the coreference of phenomenal and physical concepts. In fact the situation is the opposite: in *every* other case of concepts satisfying (i) and (ii), they have distinct referents. One might suppose that *recognitionality* is doing some extra work here (thus distinguishing this case from other cases involving nonrecognitional concepts), but the only work it does for Loar is in explaining (i) and in providing cases of coreference when (ii) is false. So once (i) and (ii) are granted, there is nothing in Loar's account to justify coreference, and his explanation of strong necessity fails.

Loar slides over this point repeatedly. His argument often seems to have the form (1) *Given that* phenomenal concepts and physical-theoretical concepts corefer (as is typical with recognitional concepts), and (2) *given that* phenomenal concepts express the property that they refer to (as is unusual

with recognitional concepts), then (3) phenomenal concepts will express physical-theoretical properties. This argument is valid, but Loar fails to note that premise (2) undercuts the support for premise (1).

A final way to put the problem. Loar's account requires, and rests on the claim, that a phenomenal concept ("feels like this") might conceive of a physical property under a necessary mode of presentation. This in turn rests on the claim that a physical property might have this mode of presentation necessarily; that is, it rests on the claim that there might be a necessary connection between a physical property and a phenomenal feel. But this is just the sort of strong necessity that was in question and that we were trying to explain. So rather than explaining strong psychophysical necessities, Loar's account presupposes strong psychophysical necessities.

We see, then, that neither Hill & McLaughlin's nor Loar's account can explain strong psychophysical necessities. Instead, both accounts *assume* such necessities at a key point, as a kind of primitive. Such necessities remain as mysterious and problematic as ever.

3.5 Modal rationalism

Hill & McLaughlin suggest that defeating all other candidates for strong necessities doesn't count for much; all that it does is remove a few "counterexamples" to the two-dimensional analysis, or to the conceivability/possibility thesis. I think this analysis gets the dialectic wrong: most type-B materialists (even Hill 1991) support their case by appealing to these examples; and the reason that many are skeptical that conceivability implies possibility is that they think there are clear counterexamples to the thesis. So removing these counterexamples removes the central plank of support from the type-B position. And the failure of any apparent explanation of strong necessities weakens the position further.

Still, the position remains at least formally open. An opponent *could* hold that there are strong psychophysical necessities quite unlike the usual *a posteriori* necessities, that the space of metaphysically possible worlds is smaller than the space of logically possible (or conceivable) worlds, and that zombie worlds are excluded. If my arguments have been successful so far, I have removed many of the positive reasons for believing this. But perhaps Hill and McLaughlin's point is that I have not shown why it *couldn't* be true. So, why must it be false?

That is a deep and interesting question on which I hope to write at length in the future, but here I will say just a few words. Of course there are the further considerations against strong necessities that I raise in the book: they they will be brute and inexplicable, that they lead to an *ad hoc* proliferation of modalities, and that they will be epistemically just like fundamental laws. And their introduction would lead to a philosophical revolution far more radical than Kripke's. But an opponent might bite the bullet on all these points. What is fundamentally *wrong* with the idea?

The fundamental problem with the idea, I think, is that it rests on a false conception of modality. It is widely acknowledged that there is a circle of modal notions: possibility, consistency, and rational entailment, for example. Some of these, and perhaps the most grounded of all, are clearly *rational*

notions, such as consistency and rational entailment. The framework of possible worlds is tremendously valuable in making sense of these notions; and as I suggested earlier, the space of worlds that we need here is the space of logically possible worlds, with one world for every ideally conceivable scenario. A narrower class of worlds is no help in making sense of these notions; breaking the tie between conceivability and possibility breaks the tie between rationality and modality.

Some might think that we cannot "break out of the circle" from rational notions to such "metaphysical" matters as *a posteriori* necessity and the concept/property distinction, but that is wrong. The point of the two-dimensional analysis of *a posteriori* necessity is that it allows us to explain such matters with only one modal primitive (the space of logically possible worlds). Given this primitive, we need only the two-dimensional intensional structure of our concepts (an *a priori* matter of conceptual analysis) along with nonmodal facts about the actual world to ground all these matters of metaphysical modality.

So one modal primitive (plus conceptual analysis plus nonmodal fundamentals) gives us everything. And it must be a primitive tied constitutively to such rational notions as consistency, entailment, and ideal conceivability. For if it is not tied constitutively to those notions, we will not be able to explain those notions (nonmodal facts certainly won't help us here), and such notions are part of what a theory of modality needs to handle. If our choice of primitive is a space of worlds, it is clearly the logically possible worlds that we need; and if our choice of primitive is not a space of worlds, it will be some other notion in the rational/modal circle.

Advocate of strong necessities must reject this picture. They cannot reject a rational modality altogether, as they use such modal notions as consistency, rational entailment, and conceivability themselves (their position is partly defined with respect to these notions). So they must accept something akin to the space of logically possible worlds, although they might use another name. But they think there is a further metaphysical modality, and that not every logically possible world is a metaphysically possible world. This modality is not to be defined in terms of the rational modality, or even in terms of the rational modality plus nonmodal facts; it is a further primitive, part of a different circle entirely.

This picture is *modal dualism*. It requires two modal primitives: there is a space of logically possible worlds, and then (as a further primitive fact) a smaller space of these picked out as "metaphysically possible". The primitives need not be cast in these terms, but we know that two primitives will be needed, as we need one to account for the rational modalities, and we need another in turn to account for the "metaphysical".

Once we get this far, it is clear that something has gone wrong. There is no reason to believe in any more than one modal primitive, and there is no reason to postulate a second "metaphysical" primitive at all. The second primitive is an *invention*; nothing in our conceptual system requires it. In particular, we need it to account for none of our existing modal notions. So it is a primitive that answers to no-one and does no work.

It seems to me that we do not even have a distinct *concept* of metaphysical necessity to which the second primitive can answer. The momentary impression of such a concept may be a residue from initial

impressions of the Kripkean distinction between epistemic and metaphysical modality. But once we recognize that this distinction can be explained with one modal primitive, and that there are constitutive ties between the Kripkean modalities, the grounds for this impression disappears. The only concept of a "metaphysical possible world" that we have is that of a logically possible world. If someone thinks they have a distinct concept here, there is no reason to believe that anything answers to it.[*]

*[[One sometimes hears phrases such as "ways the world *really* could have been", or "worlds that God could have created" as a way of trying to draw a distinction, but I cannot see any real content in these phrases over and above the invocation of logical possibility. (Of course a theist could take the second phrase literally, and perhaps call the resulting modality "metaphysical necessity". This way we would use God to ground a modal dualism. Even so, it's not clear why God's powers should prevent him from creating any logically possible world, except perhaps for those worlds not created by a God. So it may be that only strong necessity that this modal dualism would support is the existence of God!)]]

I think there is overwhelming reason to reject modal dualism and accept only one modal primitive. If so, strong necessities are ruled out, and type-B materialism is false.

3.6 Type-B materialism: Loose ends

(1) Hill & McLaughlin say at one point that the counterintuitive nature of my conclusion is itself a flaw in my treatment of modality. I think this is making the counterintuitivity do double duty. To be sure, the conclusions give reasons to resist the argument, and to look long and hard at it, but they are doing that already. But the argument itself must be assessed on its merits. Hill & McLaughlin also suggest that my view is "wishful thinking", but I think the situation is precisely the opposite. If it were not that the antecedent impulse to believe materialism were so strong (I share it, too), and my conclusions so hard to accept, I think the arguments would be relatively uncontroversial. As things stand we are in the situation where we scratch around for a materialist "way out", and the existence of any potential loophole in the argument, no matter how *ad hoc*, is seized upon. I don't say this is a bad strategy, given one's initial subjective probabilities, but I think it is clear where the wishful thinking lies.

(2) Loar mentions the explanatory gap. I think this is indeed a problem for the type-B materialist view (how *could* a physical process necessitate conscious experience? why *couldn't* God have created a zombie world?) and I think Loar's answer doesn't work, for the reasons I've given. But even if a type-B materialism is accepted, the explanatory gap is still a major issue. Loar treats the explanatory gap as if the only issue is whether it threatens materialism, and here he is like a lot of philosophers who take the truth of materialism to be the major question in this whole area; but this is not my view. There are questions about explanation, and there are questions about ontology, and the former are as at least as important as the latter. In the book, I devote at least as much space to the explanatory issues, and in the end they may matter more in developing a theory of consciousness.

Although a type-B materialism will reject my conclusions about ontology, it will (in effect) accept my conclusions about explanation. It remains the case that crossing the gap requires epistemically primitive bridging principles. These principles will be called "identities" or "necessities" rather than "laws", but their role in a theory will be much the same. A theory of consciousness will still need to have a primitive

vertical structure, not derivable from the horizontal structure of physical theory. No matter how we interpret the ontology, the shape of a theory of consciousness will be entirely different from theories in other domains.

So even if one saves the letter of materialism, it is not clear that one saves the spirit. The materialist dream of a seamless explanatory web from physics on up will fail, and we will need to search for a theoretically independent bridge. We will eventually want to systematize and simplify this bridge until we are left with a set of simple (fundamental?) identities or necessities from which all the others follow. The result will be something much like the sort of psychophysical theory that I advocate in my book (and the specific suggestions I make in the second half of the book may apply equally well or badly here). In a way, the type-B materialist ought to be as concerned with the search for a "fundamental theory" as I am.

4 Epiphenomenalism and the paradox of phenomenal judgment

Hill & McLaughlin say that I endorse epiphenomenalism, and that my anti-materialist argument implies epiphenomenalism. This is not strictly true. In fact perhaps my favorite position on the mind-body problem (as Yablo in effect notes) is not epiphenomenalism but the "panprotopsychoist" (or "Russellian") position on which basic physical dispositions are grounded in basic phenomenal or protophenomenal properties. Far from making experience causally irrelevant, this view holds that experience will be part of the categorical grounds of causation.

I think there is also more to say about interactionism than I said in my book. In particular, I think there is no knockdown argument against a quantum interactionism on which consciousness is the categorical basis of wave-function collapse. I don't favor this view myself, but it needs to be taken seriously.

Both the panprotopsychoist view and the quantum interactionist views are counterintuitive, but both are elegant and appealing and not obviously false. In both cases there are questions about whether they can really be made to work, but those issues remain open. So the conclusion of my anti-materialist argument ought not to be taken as epiphenomenalism. Rather, it is the disjunction of panprotopsychoism, epiphenomenalism, and interactionism. The question between these three views is really a distinct issue, as it can only be decided by considerations further to the argument against materialism. For my part, I rank the options in roughly the order given, but it is a deep question on which I have no settled opinion.

On none of these positions is it true that we need "an independent swarming mass of fundamental psychophysical laws", as Hill & McLaughlin suggest. We need fundamental psychophysical laws (or in the panprotopsychoist case, we need principles associating physical dispositions with (proto)phenomenal bases), but there might be very few of them, perhaps only one. (Or two, in the interactionist case: one in each direction.) It is not unreasonable to suppose that there might be a single basic psychophysical principle (some sort of transformation operator, for example) that when applied to a specific complex physical system yields a specific complex phenomenal state.

It is true that I say that phenomenal states are in a sense explanatorily irrelevant to behavior; and I think

that remains true on each of these views, as well as on the type-B materialist view. In each case, the very conceptual possibility of varying phenomenal states while preserving patterns of causal interaction suggests that explanations of behavior can be given in a way that makes no deep appeal to the phenomenal states in question. But that is not to say that they are *causally* irrelevant, except on the epiphenomenalist view.

I note also that Hill & McLaughlin are wrong in suggesting that I think sensory states are explanatorily irrelevant to our beliefs that we are in them. I do say that they are explanatorily irrelevant to the *judgments* we make about them, but this is in a stipulative functional sense of judgment, on which it comes to not much more than a disposition to report. I am quite clear, though, that experiences are not explanatorily irrelevant to our *beliefs* about them, and in fact often directly constitute aspects of those beliefs (see especially the last section of Chapter 5).

Shoemaker is right that my response to the paradox of phenomenal judgment is one of the weaker parts of the book. If I were writing Chapter 5 now, I would stress much more strongly the feature I have just mentioned: the role that phenomenal states play in constituting our phenomenal beliefs. While I noted this feature in the book, I tried to give an answer to the paradox that did not rely on it. I now think this was a mistake, as the constitutive relation between phenomenal states and associated phenomenal beliefs is crucial to the epistemology of phenomenal belief. (In effect, I would expand on the points in the last section of the chapter, particularly the parenthetical suggestion on pp. 207-8.) It is true that the paradox still applies to (functionally individuated) judgments and reports; but it is beliefs that are crucial in epistemology.

Shoemaker raises two very interesting issues related to the paradox. First, he suggests that a *zombie's* phenomenal judgments will be causally related to physical states in a way that is sufficient for reference and knowledge; so a zombie's judgments will refer to those states and will express knowledge of them; so my phenomenal judgments refer to physical states, too, as everything going on in the zombie is going on in me.

I deny the first step here. I think that for phenomenal concepts, no mere causal relation suffices for reference; and the same goes for a zombie's analog of a phenomenal concept. As I say in the book, the causal theory of reference is by no means universal. In particular, the concepts to which it applies are generally *causal concepts*; at some level, their referents are conceived as the cause of something or other, if only implicitly. This is reflected in the fact that for most of our concepts (e.g. "atom" or "dog"), the notion that the referent of the concept is wholly epiphenomenal is epistemically impossible (it is epistemically impossible that all atoms or dogs are epiphenomenal). But phenomenal concepts are not causal concepts, as witnessed by the fact that epiphenomenalism about experience is epistemically possible, if counterintuitive. The same goes for a zombie's analog of a phenomenal concept. So there is no reason to believe that a causal relation suffices for reference here, and in fact there is good reason to believe that a zombie's phenomenal concepts do not refer. (In the zombie world, zombie eliminativists are right!)

In his ingenious second point, Shoemaker notes that on my account, a zombie's phenomenal judgments

are false and unjustified, but the zombie does not seem to display any failure of rationality. But if the zombie's judgment is both rational and unjustified, then the concept of justification is cut loose from its rational moorings. Thus a conceptual incoherence seems to threaten. This is a potential problem for any view on which zombies are conceptually possible, including not only my view but also those of Hill & McLaughlin, Loar, and Yablo. What should we say?

I think the best thing to say is that the zombie's judgment is (a) unjustified (or at least not as justified as a conscious being's corresponding judgment), but (b) at least negatively rational in the sense that the zombie makes no invalid inferences and other such errors of reasoning. Justification and negative rationality often go together, but not always; this is a case where justification and negative rationality come apart. (Other such cases might be provided by an externalist view of justification, where justification is not always grounded in rationality. This is not to say that the zombie case involves externalist justification: my analysis holds onto the internalist idea that justification is a matter of having evidence.) I think it is plausible that where truly basic evidence is concerned, as with the relation between an experience and the associated phenomenal belief, the justification relation is something tighter than any cognitive relation involving mere negative rationality. This fits with the idea mentioned above, that the justification comes from the constitutive relation between experience and belief. If this is so, we should not expect the justification to be a wholly rational matter.

The zombie lacks evidence for its judgment (so it is unjustified), but it does not know that it lacks this evidence, and it is not culpable for this lack of knowledge (it is negatively rational). Shoemaker's "rational moorings" for justification may still be preserved in a link between justification and positive rationality, which requires sufficient reasons and evidence for a judgment. But justification (especially basic justification) and positive rationality involve more than a mere absence of mistakes, so negative rationality does not in general suffice.

One could also take a more extreme line, and deny that notions such as rationality and justification apply to zombies at all. I have gradually become more sympathetic with the idea that phenomenology plays a role in constituting intentionality, so that a zombie has intentional states in at best a weak sense. But I do not want to rely on that point here.

5 Dancing qualia

Shoemaker also comments on the dancing qualia case. I think he is certainly right that qualitative beliefs will change when qualia change. This follows from the constitutive relation between experience and belief mentioned above; and it is how things *must* go, in order that the post-flip state not be incoherent. I did not mean to deny that beliefs would change in this sort of way. I was instead focusing on the absence of a belief *about the change*.

The surprising part of the dancing qualia case is not that the subject (after the flip) has false beliefs about past qualia, or even that he does not introspectively realize that his qualia are different. These things hold for any case of inverted qualia. The surprising part of the dancing qualia case is that the flip should go

unnoticed when it happens. To have one's qualia change massively before one's eyes will be a major phenomenal event; but it will be a phenomenal event to which we have no access. Of course Shoemaker is right that *if* inverted qualia are possible, we should *expect* unnoticed changes in this scenario (because beliefs and memories will change too). But I don't think this diminishes the implausibility of the consequent here; rather, it indirectly adds to the implausibility of the antecedent.

Shoemaker thinks the surprisingness can be explained away in this case, because there is an "unnatural tampering with cognitive dynamics". I am not sure that it is all that unnatural. In future years, we might end up performing this sort of substitution with some regularity; and on a lower level, physiological changes with little impact on function may already be happening all the time. It's true that if we assume that qualia are changing massively, then there is a sense in which the change in "unnatural", but the question of whether qualia change massively is precisely what is at issue. Characterized in neutral terms, the change to one's underlying processes does not seem especially unnatural; so it is surprising that it should produce such a massive, unnoticed change.

In any case the dancing qualia case was never intended as a knockdown argument, and there are certainly ways in which it can be resisted. On balance of intuition, I have come to think that in some ways the fading qualia argument is stronger than the dancing qualia argument: the idea of a system continually and completely out of touch with its "faded" qualia is even stranger than the idea of a system momentarily out of touch with its massively changing qualia. (Of course this is a conclusion Shoemaker would be happy with, as he thinks that absent and fading qualia are logically impossible, while inverted and dancing qualia may be naturally possible.) But at the end of the day these arguments are really intended as thought-experiments that any theory must face up to and make sense out of. Proponents of a different view might bite the bullet on any of them, as long as they are aware of the costs. I think that other things being equal, the conclusion that qualia do not fade and dance unnoticed is more plausible than the alternatives, but as with all of these matters, there is probably more to be said.[*]

*[[Thanks to Tamar Gendler, Matthew Henken, David Lewis, Gregg Rosenberg, and Stephen Yablo for comments.]]

References

- Hill, C.S. 1991. *Sensations: A Defense of Type Materialism*. Cambridge University Press.
- Hill, C.S. 1997. Imaginability, conceivability, possibility, and the mind-body problem. *Philosophical Studies* 87:61-85.
- Hill, C.S. 1998. Chalmers on the apriority of modal knowledge. *Analysis* 58:20-26.
- Loar, B. 1997. Phenomenal states (second version). In (N. Block, O. Flanagan, and G. Güzeldere, eds) *The Nature of Consciousness: Philosophical Debates*. MIT Press.

Precis of *The Conscious Mind*

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

Chapter 1: *Two Concepts of Mind*. I distinguish the phenomenal and psychological (functional) concepts of mind. I argue that every mental state is a phenomenal state, a psychological state, or a hybrid of the two. I discuss the two mind-body problems corresponding to the two concepts of mind, and discuss the various senses of the term "consciousness".

Chapter 2: *Supervenience and Explanation*. I distinguish varieties of supervenience, especially logical and natural supervenience, where supervening properties covary with base properties across either logically possible (conceivable) or naturally possible worlds. (One can also define a distinct notion of metaphysical supervenience, if one believes that the classes of logically possible and metaphysically possible worlds are distinct; I hold that the two classes coincide.) *Materialism* is the thesis that all positive properties globally metaphysically supervene on microphysical properties. A property is *reductively explainable* in terms of physical properties if it globally logically supervenes on those properties.

I defend the notion of conceptual truth against various objections, and outline a two-dimensional intensional framework for handling *a posteriori* necessity. Every concept has two intensions, a primary intension and a secondary intension. The primary intension delivers a concept's referent in a centered world when the world is considered as actual (i.e., considered as an epistemic possibility); the secondary intension delivers a concept's referent in a world when it is considered as counterfactual. (The primary intension of "water" picks out the "watery stuff" in a world; the secondary intension picks out H₂O.) A statement is *a priori* when it has a necessary primary intension; a statement is necessary when it has a necessary secondary intensions. *A posteriori* necessities (such as "water is H₂O") have a necessary secondary intension but a contingent primary intension. A statement is conceivable (or logically possible) when its primary intension is true in some world; a statement is possible (metaphysically possible) when its secondary intension is true in some world. So the Kripkean gap between conceivability and possibility is explained at the level of statements, without appealing to a distinction between conceivable and possible worlds. The class of worlds in question is always that of the ideally conceivable (or logically possible) worlds.

I use arguments from conceivability, epistemology, and analysis to argue that *almost* all properties logically supervene on the physical, except perhaps for phenomenal properties and other properties depending on them.

Chapter 3: *Can Consciousness be Reductively Explained?*. I use arguments from conceivability, epistemology, and analysis to argue that phenomenal properties do not logically supervene on (are not *a priori* entailed by) physical properties. These include arguments from the conceivability of zombies and inverted spectra, arguments from incomplete knowledge and from epistemic asymmetry, and arguments from the absence of structural/functional analysis. It follows that there is no reductive explanation of consciousness. I apply this to potential explanations in neuroscience, cognitive science, and physics.

Chapter 4: *Naturalistic Dualism*. I argue that phenomenal properties are not necessitated by physical properties, so that materialism is false. The arguments of Chapter 3 are supplemented by an argument that there is no *a posteriori* necessary entailment from physical to phenomenal. I use the two-dimensional analysis of *a posteriori* necessity to argue that no such necessity can save materialism if *a priori* entailment fails. And I argue that there is no reason to believe in "strong" metaphysical necessities that escape the two-dimensional account. I also discuss and argue against the possibility that there is an *a priori* connection that we cannot grasp due to cognitive limitations.

I compare this argument to related arguments for dualism and discuss the position that results. Materialist ontology must be expanded by new fundamental properties (phenomenal or protophenomenal properties) and new fundamental psychophysical laws. The resulting position may be epiphenomenalism, interactionism (although I argue against this), or a Russellian monism on which (proto)phenomenal properties serve as the categorical grounds of basic physical dispositions. I sketch the logical geography of the issues and defend my "naturalistic dualism" from objections.

Chapter 5: *The Paradox of Phenomenal Judgment*. On my position, even if consciousness cannot be physically explained, behavior and functioning can be. So it seems that consciousness is explanatorily (although perhaps not causally) irrelevant to behavior. In particular it is explanatorily irrelevant to claims such as "I am conscious" and related phenomenal judgments (where judgments are defined in functional terms). I call this the "paradox of phenomenal judgment". I argue that this paradox is counterintuitive but poses no fatal flaws. I address the objections that it implies that we are unable to know about, refer to, or remember our phenomenal states. I argue that these objections rest on causal theories of knowledge and of reference that we have independent reason to reject in the phenomenal case. Knowledge of and reference to phenomenal states is based on something tighter than a causal relation; it is based on a relation of acquaintance. I discuss the content of phenomenal beliefs and the constitutive relation between experience and phenomenal belief.

Chapter 6: *On the Coherence between Consciousness and Cognition*. I outline the program of searching for a theory of consciousness in terms of psychophysical laws: first nonfundamental, then fundamental laws. Data is hard to come by, but first-person data plus plausibility assumptions in the third-person case can do much work. I discuss some candidates for nonfundamental psychophysical laws, involving the "coherence" between consciousness and "awareness", a functional state defined in terms of direct availability for global control. I argue that consciousness and awareness covary, as do their structures. I apply this to some methodological questions about the empirical study of consciousness.

Chapter 7: *Absent Qualia, Fading Qualia, Dancing Qualia*. I argue for a principle of organizational invariance: systems with the same fine-grained functional organization will have the same sort of phenomenal states (with natural necessity). I argue against the empirical possibility of "absent" and "inverted" qualia, using thought-experiments involving replacement of neurons by silicon chips. If absent qualia are possible, then "fading" qualia are possible, and if absent or inverted qualia are possible, then "dancing" qualia are possible; but I argue that it is very implausible that fading or dancing qualia are possible.

Chapter 8: *Consciousness and Information: Some Speculation*. I speculate on the form of a fundamental theory of consciousness, suggesting that it may involve a dual-aspect view of information (information realized both physically and phenomenally). I argue that this would cohere well with the other principles, and with the relation between consciousness and phenomenal claims. One version of this view implies a sort of panpsychism; this need not follow, but I argue that it is not as implausible as is often supposed. I suggest that such a metaphysics of information may fit well with the Russellian monism discussed earlier, and discuss some open questions for such a theory.

Chapter 9: *Strong Artificial Intelligence*. I apply previous conclusions to argue that strong artificial intelligence is true: there is a class of programs such that any implementation of a program in that class is conscious. I give an account of computation and implementation to support this conclusion, and I argue against Searle's "Chinese room" arguments and other objections to strong artificial intelligence.

Chapter 10: *The Interpretation of Quantum Mechanics*. I discuss some relations between a theory of consciousness and the interpretation of quantum mechanics. After discussing the measurement problem and various interpretations, I defend a version of the Everett interpretation on which wavefunctions never

collapse. A major question for this view is why, given that the physical world is superposed, we experience it as discrete. I suggest that this is a question for a theory of consciousness, and argue that the principles I have developed *predict* that there will be discrete experiences in a superposed world. I defend the Everett interpretation from some objections, although I am ultimately agnostic about its truth.

MIND AND MODALITY

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

What follows are compressed versions of three lectures on the subject of "Mind and Modality", given at Princeton University the week of October 12-16, 1998. The first two form a series; the third stands alone to some extent. All are philosophically technical, and probably of interest mainly to philosophers. I hope that they make sense, at least to those familiar with my book *The Conscious Mind*.

Lecture 1 recapitulates some of the material in the book in a somewhat different form, and adds some further material on conditionals and on Kripke. Note that [section 5](#) has a more or less definitive formalization of the anti-materialist argument from the book (lots of people have asked for this). Lecture 2 pushes deeper into the heart of modality, further investigating the conceivability/possibility relationship and the epistemology of modality (with some material on the scrutability of truth in general), and arguing for a sort of modal rationalism. Lecture 3 gives an analysis of the content of beliefs about experiences, and applies this to a number of epistemological issues, including incorrigibility and the dialectic on "The Myth of the Given".

Much of this material is being further developed in work in progress, so comments are very welcome. The material in the first two lectures may see the light of day either as a series of papers or as a book. For an earlier, non-outline version of a few of these ideas, see my response (["Materialism and the Metaphysics of Modality"](#)) in the *Philosophy and Phenomenological Research* symposium on my book. Many of the ideas in lecture two are elaborated in much more detail in my newer (summer 1999) paper ["Does Conceivability Entail Possibility?"](#). The ideas in lecture three now form the basis for a separate long paper, ["The Content and Epistemology of Phenomenal Belief"](#).

My Arizona graduate seminar (spring 1999) on "Mind and Modality" largely followed the content of these lectures. The [web page](#) for that seminar has links to a lot of e-mail discussion in which I and others discussed many of these ideas in more depth.

-
- Lecture 1: [Consciousness, Conceivability, and Conditionals](#)
 - Lecture 2: [Modal Rationalism and the Scrutability of Truth](#)
 - Lecture 3: [The Content and Epistemology of Phenomenal Belief](#)
-

1: CONSCIOUSNESS, CONCEIVABILITY, AND CONDITIONALS

(1) EPISTEMIC ARGUMENTS AGAINST MATERIALISM

The phenomenal facts are not entailed *a priori* by the physical facts. (See *The Conscious Mind*, chapter 3.)

- (a) Conceivability of zombies
- (b) Conceivability of inverted spectra and other phenomenal variants
- (c) Knowledge argument
- (d) Epistemic asymmetry
- (e) Absence of structural/functional analysis of phenomenal concepts

(2) MATERIALIST RESPONSES

(i) **Type-A materialism:** Deny the epistemic intuitions, and assert *a priori* entailment (eliminativism, analytic functionalism, ...). Not considered here.

(ii) **Type-B materialism:** Accept the epistemic intuitions, but hold that there is an *a posteriori* necessary entailment from physical to phenomenal. Physical:phenomenal as water:H₂O or Hesperus:Phosphorus. Conceivability doesn't imply possibility.

(3) TWO-DIMENSIONAL ANALYSIS OF A POSTERIORI NECESSITY

This framework analyzes *a posteriori* necessity in terms of two-dimensional semantic evaluation over one space of possible worlds. Every statement S (as used by a given speaker on a given occasion) has two intensions, a primary and secondary intension. (See *The Conscious Mind*, section 2.4, and [The Components of Content](#).)

Primary intension of S: (or **epistemic profile** of S). Evaluate at a centered world W.

W → truth-value of S in W when W is considered as actual (i.e., considered the way we consider epistemic possibilities).

Heuristic: If W is the case, is S the case? [Indicative conditional.]

- "water is H₂O": the set of worlds in which watery stuff is H₂O-structured
- "Hesperus is Phosphorus": worlds in which morning star is evening star

Secondary intension of S: (or **modal profile** of S). Evaluate at an uncentered world W.

$W \rightarrow$ truth-value of S in W when W is considered as counterfactual.

Heuristic: If W were the case, would S be the case? [Subjunctive conditional.]

- "water is H₂O": the set of worlds in which H₂O is H₂O
- "Hesperus is Phosphorus": the set of worlds in which Venus is Venus

S is "necessary" $\rightarrow$ S has a necessary secondary intension.

S is *a priori* $\rightarrow$ S has a necessary primary intension

S is *a posteriori* necessary $\rightarrow$ S has a contingent primary intension and a necessary secondary intension.

Supports *modal rationalism*: when S is conceivable (approx: not-S is not *a priori*), S describes a possible world, at least according to its primary intension (although perhaps not according to its secondary intension). We retain *a priori* access to the space of possible worlds.

(4) DIVERSION: THE TYRANNY OF THE SUBJUNCTIVE?

An important difference between indicative and subjunctive conditionals:

(1a) "If Prince Albert killed those people, he is Jack the Ripper."

(1b) "If Prince Albert had killed those people, Jack the Ripper wouldn't have".

(2a) "If XYZ is the liquid in the oceans and lakes, water is XYZ."

(2b) "If XYZ were the liquid in the oceans and lakes, water wouldn't be in the oceans and lakes".

These conditionals give what is at least the intuitively correct thing to say about the consequent, given the antecedent. General moral: (At least on an intuitive reading) *indicatives consider the antecedent as actual; subjunctives consider the antecedent as counterfactual*. Indicatives track primary intensions; subjunctives track secondary intensions.

The analysis of necessity and possibility:

S is necessary $\leftrightarrow$ S is true in all possible worlds.

S is possible $\leftrightarrow$ S is true in some possible world.

QUESTION: How to evaluate "S is true in W"?

SUBJUNCTIVE READING: "S is true in W" iff

If W were the case, S would be the case.

INDICATIVE READING: "S is true in W" iff

If W is the case, S is the case.

There are two resulting sorts of "necessity": *subjunctive necessity* and *indicative necessity*. And two sorts of "possibility", etc. Indicative necessity is necessity according to primary intensions; subjunctive necessity is necessity according to secondary intensions.

In contemporary philosophy (at least since Kripke), "necessity" (simpliciter) is always read as subjunctive necessity, and "P is true in W" is always evaluated by the subjunctive reading.

In fact: subjunctive conditionals *ground* the contemporary evaluation of statements in worlds, and the consequent analysis of necessity. Kripke's ultimate court of appeals in *Naming and Necessity* is always a subjunctive.

QUESTIONS: Why is this? Is it an arbitrary choice? Is the choice argued or stipulated?

One might propose a **SYMMETRY THESIS**: Indicative necessity is as good a candidate for "necessity" as subjunctive necessity. Is this viable?

[If so, it would be just as reasonable to say that "Hesperus is visible in the evening (if it exists)" is necessary; that "water is H₂O" is contingent; that "I am here now (if I exist)" is necessary; that "the meter stick in Paris is a meter long (if it exists)" is necessary; that the necessity of identity doesn't hold; etc.

Imagine an alternative universe in which Kripke used indicatives instead of subjunctives. The resulting book, *Naming and Non-Necessity*, might have defended a strong link between the a priori and the necessary, have argued against the necessity of identity and *de re* necessities, and so on. Direct reference theory, singular propositions, and externalism have a much harder time getting off the ground. The theory of content that ensues has much less trouble meshing with a theory of cognition. And so on.

Or imagine a symmetrical universe, a history in which the notion of necessity is seen to divide into two sorts, indicative and subjunctive necessity. The former tracks epistemic notions more closely than the latter. One can do quantified modal logic with the latter but not the former (thus a split decision in the debate between the likes of Quine and the likes of Kripke)? Whenever one speaks of truth in a possible worlds, one has to be explicit about the means of evaluation, rather than presupposing the subjunctive evaluation. Subjunctive necessity gives a nice metaphysics of essentialism and a semantics for counterfactuals, but indicative necessity yields a more powerful theory of content. A more balanced 30 years of philosophy?]

SIX POSSIBLE REASONS FOR FAVORING THE SUBJUNCTIVE

- (a) "Could not have been otherwise".
- (b) Indicative necessity is "merely epistemic".

- (c) Indicative necessity requires centered worlds.
- (d) Subjunctive necessity supports quantified modal logic, etc.
- (e) The relevant indicative conditionals aren't true/false, just assertible, due to epistemic relativity.
- (f) Indicative "necessities" will vary in truth-value between users, since primary intensions can vary.

[The first four clearly don't carry much weight against the symmetry thesis. The last two carry a little, but not an overwhelming amount. The epistemic relativity in (e) applies only to incomplete antecedents - how one fleshes out an incomplete antecedent into a world depends on what one believes - but the underlying evaluation in a world is unaffected. The variability in (f) is present in reverse, for secondary intensions and subjunctive necessities, with terms such as indexicals and demonstratives, but that isn't taken to undermine claims of necessity of utterances here; so the same should arguably go for indicative necessity. So the symmetry thesis is still tempting.]

(5) APPLICATION OF 2-D ANALYSIS TO THE ARGUMENT AGAINST MATERIALISM

[This is a crisper version of the argument in *The Conscious Mind*, section 4.2]

Premise 1: The physical truths don't a priori entail the phenomenal truths.

Premise 2: When S is a posteriori, S has a contingent primary intension.

Premise 3: If materialism is true, the physical facts necessitate all true propositions.

4: $P \rightarrow Q$ is a posteriori. [P = physical truths, Q = a phenomenal truth] (from 1)

5: $P \rightarrow Q$ has a contingent primary intension (from 2, 4)

6: Exists W: P' is true in W, Q' is false in W [P', Q' = PI's of P, Q] (from 5)

7: Exists W: P is true in W, Q' is false in W [from 6; see loophole 2]

8: Materialism is false [from 3, 7; see loophole 1]

Loophole 1: Primary intensions are indexical; Q' is a centered proposition. Materialism requires only that P necessitates all true *uncentered* propositions.

Response 1: Adding "locating" information I (an identifying description picking out an individual at a time within a world) to P doesn't help. Revised premise 1: $P \& I \rightarrow Q$ is not a priori. So $(P' \text{ at } I) \rightarrow (Q' \text{ at } I)$ is contingent. And $(Q' \text{ at } I)$ is an uncentered proposition.

[N.B. We can understand the physical facts to include a "that's-all" fact without changing anything essential.]

Loophole 2: Maybe microphysical facts have different PI/SI, so P necessitates phenomenal facts but P' [or $(P' \text{ at } I)$] does not.

Response 2: This is a valid point, but when cashed out it leads to "panprotopsyism", with fundamental protophenomenal properties as the categorical basis of microphysical dispositions. An important position, but more in the spirit of dualism than materialism.

UPSHOT: *A posteriori* necessity (at least of the usual variety) can't save materialism.

Essential point: If it is ideally conceivable that there are zombies, then the primary intension of "There are zombies" describes at least one possible world, and that suffices to ensure that materialism is false. (Conceivability implies possibility at the level of worlds, if not at the level of statements.)

(6) KRIPKE'S ARGUMENT

There is a natural mapping between this argument and Kripke's argument against materialism. Some differences and correspondences:

(i) I'm arguing against upward necessitation, not type or token identity.

[Token and type identity arguably aren't required for materialism. E.g. consider the statue and the lump of clay (re token identity), or the possibility of disembodied life (re type identity). What matters is upward necessitation.]

(ii) I appeal to physical-without-mental scenarios, but not mental-without-physical scenarios.

[The possibility of mental-without-physical doesn't falsify materialism. Again, witness the possibility of disembodied life.]

(iii) Premise 2 is somewhat akin to Kripke's two strategies for explaining away apparent contingency.

[Kripke's strategy 1: "a qualitatively analogous statement would be false in a qualitatively identical epistemic situation". The "qualitatively analogous statement" is more or less a statement with the same primary intension. But the criterion is too strong.

Consider "Bill", stipulated to be a name for the quality in the center of my visual field. Then "Bill = blue" is an "apparently contingent" necessity in the usual way. But in any qualitatively identical epistemic situation, a qualitatively analogous statement is true. Moral: drop the "qualitatively identical epistemic situation" clause,

which is unnecessary (the epistemic situation plays no special role).

Consider also "Fred", a name stipulated to pick out 1 if there is at least one statement in the world, otherwise 0. Then "Fred = 1" is another "apparently contingent" necessity. But a qualitatively analogous statement will always be true. Moral: drop the idea of evaluating a statement present at the center of a given alternative world, and instead evaluate the *actual* statement in that world considered as actual. The result of these two changes will be more or less my premise 2. (N.B. The primary intension of "Fred = 1" is false in an empty world, even though there is no statement to evaluate there. Indicatively: if the empty world is actual, then Fred = 0.)

Kripke's strategy 2, "contingently coinciding reference-fixers for an identity", can be seen to be a special case of premise 2.]

(iv) Kripke's premise "pain = felt-as-pain" is not required.

[Maps to: primary intension = secondary intension for phenomenal concepts. This is plausible but unnecessary for the argument to go through.]

(v) One needs to handle primary/secondary gap for physical concepts.

[Kripke only discusses the gap, in effect, for phenomenal concepts. This gap leads to the "panprotopsychoist" escape route, discussed e.g. in Maxwell's paper on Kripke.]

(7) WHITHER TYPE-B MATERIALISM?

To resist, a type-B materialist must deny the 2-D analysis of *a posteriori* psychophysical necessities. Psychophysical necessities are different in kind from standard *a posteriori* necessities.

Must deny premise 2. Psychophysical necessities, although *a posteriori*, have a metaphysically necessary primary intension.

Must deny modal rationalism: Zombies, although conceivable, are strongly metaphysically impossible: there is no possible world satisfying even the primary intension of "There are zombies". The conceivable world corresponds to no possible world at all.

More on this in lecture 2.

(8) CONCLUSION

If the anti-materialism argument is accepted (so type-A and type-B materialism are rejected), we're left with about three positions to choose from. All might be called versions of "phenomenal fundamentalism". The conclusion of the argument should best be seen as a disjunction of these.

(i) *Epiphenomenalism*: Causally closed physical world, naturally supervenient phenomenal properties.

[No knockdown arguments against this, although counterintuitive and more importantly inelegant.]

(ii) *Interactionism*: No causal closure, two-way psychophysical laws.

[Compatible with one natural interpretation of quantum mechanics.]

(iii) *Panprotopsyism*: Fundamental protophenomenal properties serve as the categorical basis for microphysical dispositions in a causally closed physical world.

[Might be seen as materialism, or as non-Berkeleyan idealism, or (perhaps best) as a dual-aspect view with an underlying neutral monism. Deeply attractive, if the constitution relations can be figured out.]

2: MODAL RATIONALISM AND THE SCRUTABILITY OF TRUTH

[Much of the material in this lecture is elaborated in detail in my paper ["Does Conceivability Entail Possibility"](#).]

(1) YESTERDAY

(a) Epistemic arguments against materialism. The phenomenal facts are not entailed a priori by the physical facts. (Zombies, inverted spectra, knowledge argument, etc.)

(b) Type-B materialism: There is an a posteriori necessary entailment from physical to phenomenal.

(c) Two-dimensional analysis of a posteriori necessity. Every statement S (as used by an individual) has two intensions, a primary and a secondary intension. Primary intension evaluates S in a (centered) world when the world is considered as actual; secondary intension evaluates S in a world when the world is considered as counterfactual. S is "necessary" when it has a necessary secondary intension. If S is a priori, it has a necessary primary intension. A posteriori necessities have contingent primary intension and necessary secondary intension.

(d) Argument against materialism. Premise 1: The physical facts don't a priori entail the phenomenal facts. Premise 2: An a posteriori statement has a contingent primary intension. Conclusion: Materialism is false.

(e) To resist: a type-B materialist must deny premise 2, and deny the 2-D analysis of a posteriori

psychophysical necessities. These necessities, unlike standard a posteriori necessities, do not have a contingent primary intension. Must deny modal rationalism: Zombies, although conceivable, are strongly metaphysically impossible: there is no possible world satisfying even the primary intension of "There are zombies". The conceivable world corresponds to no possible world at all.

(2) STRONG NECESSITIES

Crucial issue: Are there strong *a posteriori* necessities? ["Strong metaphysical necessities", in *The Conscious Mind*.]

Strong necessity = an a posteriori necessity with a metaphysically necessary primary intension.

Intuitively: If S is a posteriori, there is some conceivable "world" W such that if W is actual, S turns out to be true. If S is a standard a posteriori necessity, W is possible. If S is a strong necessity, W is not possible.

(3) DOES CONCEIVABILITY IMPLY POSSIBILITY?

Distinguish three independent dimensions in notions of conceivability (of a statement S):

Positive vs. **negative** conceivability (clear and distinct conceivability of a situation verifying S, vs. absence of any apparent contradiction in S).

Prima facie vs. **ideal** conceivability (conceivability on first appearances, vs. on ideal rational reflection).

Primary vs. **secondary** conceivability (conceivability when considered as actual vs. when considered as counterfactual).

Standard worries about conceivability vs. possibility, translated:

(1) Prima facie conceivability is an imperfect guide to possibility (but mainly insofar as it's an imperfect guide to ideal conceivability).

(2) Secondary conceivability is a posteriori (so set it aside).

(3) Primary conceivability is a poor guide to standard (secondary) possibility of S (but may be a good guide to the primary possibility of S, i.e. the possibility of a world satisfying S's primary intension).

So ideal primary conceivability is the sort that counts.

THESIS: ideal primary positive conceivability of S implies primary possibility (i.e., possibility of a world satisfying S's primary intension).

And just maybe: ideal primary negative conceivability of S (i.e., $\sim S$ is not a priori) implies primary possibility.

["Ideal primary" is tacit henceforth.]

(4) VARIETIES OF MODAL RATIONALISM

S is *a priori* if it is rationally knowable (justifiable independently of experience).

S is *a posteriori* if there is a positively conceivable scenario (considered as actual) which verifies $\sim S$.

S is in the *twilight zone* if it is neither a priori nor a posteriori (as above): not rationally knowable, but not positively conceivable otherwise. i.e., $\sim S$ is negatively conceivable, but not positively conceivable.

Deep issue: are there any twilight-zone truths? i.e., is there a gap between positive conceivability and negative conceivability?

Candidates (more on these below):

- (i) Certain mathematical statements, e.g. continuum hypothesis?
- (ii) Conditionals about (e.g.) tallness, on epistemic view of vagueness?

STRONG MODAL RATIONALISM: S is a priori iff it has a necessary primary intension

or: if S is negatively conceivable, it is possible.

WEAK MODAL RATIONALISM: If S is a posteriori, it has a contingent primary intension

or: if S is positively conceivable, it is possible.

NEGPOS PRINCIPLE: If S is negatively conceivable, it is positively conceivable.

Weak modal rationalism denies the existence of strong necessities. The negpos principle denies the existence of twilight-zone truths. Strong modal rationalism (roughly) denies both.

Here, I'll argue only for weak modal rationalism, not strong modal rationalism (though I am inclined to accept both). What matters for the mind-body issue are strong necessities, not twilight-zone truths (as zombies are positively conceivable).

(5) DIVERSION: THE SCRUTABILITY OF TRUTH

(Or: on the deep philosophical importance of the twilight zone.)

SCRUTABILITY PRINCIPLE (first pass): Once we know how the world is qualitatively, we're in a position to know what our terms refer to and whether our statements are true.

SCRUTABILITY OF REFERENCE is problematic due to (i) unclarity of "know what terms refer to", (ii) indeterminacy due to varying reference while preserving truth-value across worlds. (cf. Quine, Benacerraf, Putnam, etc). So work with truth.

SCRUTABILITY OF TRUTH (second pass): A complete enough qualitative description of the world implies all true statements. ["A implies B" means " $A \rightarrow B$ " is a priori" throughout.]

QUESTION: What counts as a "complete enough qualitative description"?

- (i) Explicitly specifies all the facts? No, that makes the claim trivial.
- (ii) A description in terms of fundamental natural properties? Not bad, but will be denied by a type-B materialist. I want to keep issue of scrutability separate from the issue of strong necessities.

But note: point of fundamental-property description is to specify the world uniquely within the space of metaphysically possible worlds. So a "complete enough qualitative description" should specify a unique positively conceivable scenario.

D is **PC-complete** if (i) D is positively conceivable; (ii) if E is positively conceivable and E implies D, D implies E.

[For a type-A materialist, a complete microphysical description of the world will be PC-complete. For a type-B materialist or a phenomenal fundamentalist, a complete physical and phenomenal description of the world will probably be PC-complete.]

SCRUTABILITY OF TRUTH: If D is a PC-complete truth and S is true, $D \rightarrow S$ is a priori.

S is an **inscrutable truth** if S is true and some PC-complete truth D doesn't imply S. [N.B. If S is inscrutable, $D \rightarrow S$ is in the twilight zone.]

Deep question: are there any inscrutable truths? Candidates:

- (i) Some mathematical statements, e.g. continuum hypothesis?
- (ii) Statements about e.g. tallness, on epistemic theory of vagueness?

[But: one can argue that all mathematical truths are a priori (our inability to know them is merely due to cognitive limitation, and disappears on idealization), and that any that are not a priori are not determinate at all. And the epistemic theory of vagueness is generally regarded as implausible.]

Generalization of scrutability: a complete qualitative description of any world (not just the actual world) leaves nothing epistemically open (except indeterminacies).

D is **NC-complete** if (i) D is negatively conceivable, (ii) if E is negatively conceivable and E implies D, then D implies E.

GENERALIZED SCRUTABILITY: If D is PC-complete, D is NC-complete.

[I am inclined to believe both scrutability and generalized scrutability, but I think this is one of the deepest issues in philosophy and requires an extended treatment in its own right. It connects not just to the theory of truth and reference but to issues of realism and anti-realism (one might think of scrutability as a watered-down "realist" version of the anti-realist "nothing is hidden" thesis: nothing is hidden given enough qualitative information). And if true it has application throughout metaphysics, ethics, philosophy of mind, etc.]

(6) DIVERSION CONTINUED: PURE MODAL RATIONALISM

S is an **open inconceivability** if S is negatively conceivable, but for all PC-complete D, D implies $\sim S$. (e.g., S = "There are inconceivable features of the world"?).

NOINCONCEIVABILITY: No S is an open inconceivability.

Then: NEGPOS $\leftrightarrow$ GENERALIZED SCRUTABILITY & NOINCONCEIVABILITY.

PURE MODAL RATIONALISM: S is positively conceivable $\leftrightarrow$ S is negatively conceivable $\leftrightarrow$ S is possible.

Pure modal rationalism $\leftrightarrow$ weak modal rationalism & negpos $\leftrightarrow$ weak modal rationalism & generalized scrutability & noinconceivability.

This requires ruling out (i) strong necessities (ii) generalized inscrutabilities (iii) open inconceivabilities. [Weak modal rationalism rules out (i), strong modal rationalism rules out (i) and (ii).]

[I'd like to believe pure modal rationalism. If true, it gives a beautifully elegant view of our epistemic access to modality. I am very confident about ruling out (i), fairly confident about ruling out (ii), and somewhat unsure about ruling out (iii), but I can at least see a case for it.]

(7) ARE THERE STRONG NECESSITIES? I. COUNTEREXAMPLES

Back to making the case for weak modal rationalism and against strong necessities, which is what's relevant to the mind-body issue.

Possible examples of strong necessities:

- (i) Standard a posteriori necessities - no good.
- (ii) Psychophysical necessities (on type-B view) - tendentious.
- (iii) Laws of nature (on some "strong law" views) - tendentious.
- (iv) God's existence (on some theist views) - tendentious.

[For more discussion see my [PPR response, section 3.3.](#)]

(8) ARE THERE STRONG NECESSITIES? II. EXPLANATIONS

In *The Conscious Mind*, I say that strong necessities will be brute facts and inexplicable. (And are thus better seen as contingent laws.) Some "explanations" have been offered in response:

(i) Hill: Explain "illusion of possibility" in psychological terms.

[Response: this doesn't explain the unreliability of modal intuition, i.e. the *co-presence* of necessity and apparent contingency; so needs to assume strong necessities as an additional part of the explanation; also, this strategy would analogously "explain away" all mathematical/modal knowledge. See [PPR response, section 3.4.](#)]

(ii) Loar: Because phenomenal concepts are recognitional concepts with primary = secondary intension.

[Response: this doesn't work unless it presupposes strong necessities to start with, as it presupposes that physical properties have a necessary phenomenal mode of presentation. Also, the recognitional argument doesn't work. See [PPR response, section 3.4.](#)]

(iii) Levine, Melnyk: We don't always have *a priori* access to primary intensions.

[Response: This is hard to support and in any case would yield inscrutable truths rather than strong necessities. See [responses to Levine and Melnyk.](#)]

(9) ARE THERE STRONG NECESSITIES? III. WHY NOT?

Materialist might say: I still haven't ruled out the possibility that strong necessities exist. Perhaps strong necessities are *sui generis* and inexplicable? Or perhaps an explanation is waiting to be found?

Q: How to argue positively for (weak) modal rationalism? (So far, we've argued negatively by defeating counterexamples and arguments against it.)

This raises the deepest philosophical issues in the vicinity. Here is an argument sketch.

Step by step...

(i) We need the modality of ideally conceivable worlds (= logically possible worlds). EVEN IF some of these worlds aren't "metaphysically possible", we need these worlds for most of the purposes that we need possible worlds in the first place.

- To make sense of counterfactual thought.
- To give an account of the contents of thought.
- To give an account of the semantics of language.
- To systematize rational inference.
- Etc.

Imagine that a "strong laws" view is true, and counternomic worlds are metaphysically impossible. (Or that a type-B materialist view is true, and zombie worlds are metaphysically impossible.) Still, we will need counternomic worlds and zombie worlds in *some* strong sense for the above purposes: to make sense of a scientist's thought processes, counterfactuals concerning counternomic scenarios, the differing meanings of nomically coextensive terms, and so on. Same for zombie worlds. (See my [PPR response, section 3.2.](#))

(ii) There is no bar to the existence of logically possible worlds. If one doesn't want to just postulate them, they are easy to construct (to "ersatz" into existence), e.g. in terms of equivalence classes of PC-complete descriptions. (With small twists to handle centering and two-dimensional semantics. The construction is an instructive enterprise.). They behave just as possible worlds should: one can tell coherent science fiction about them, one can semantically evaluate all our terms and statements in them, and one can consider them both as actual and counterfactual.

(iii) So: we need the logical modality and there is no bar to it. So insofar as there is reason to believe in modality at all, there is reason to believe in this modality.

(iv) From the logical modality, we can recover *all* modal phenomena in which we have reason to believe; and we can use it to explain everything which modality can plausibly explain. We can use it in explaining counterfactuals, the contents of thought, rational inference, the semantics of language, and so on. And with the help of two-dimensional semantics (plus nonmodal facts), we can use it to explain such "metaphysical" phenomena as *a posteriori* necessity, the concept/property distinction, and so on.

(v) So: one modal primitive (plus conceptual analysis plus nonmodal facts) gives us everything we have reason to believe in. And this modal primitive is tied constitutively to the rational notions (if it's not, we won't be able to explain them): hence, constitutive ties between rational and modal notions.

(vi) The believer in strong necessities, by contrast, must embrace a *modal dualism*: there are distinct

primitive modalities of logical and metaphysical possibility. The metaphysically possible worlds are a subset of the logically possible worlds, and neither class is reducible to the other. One modality handles "rational" modal concepts, the other handles "metaphysical" modal concepts.

(vii) There is no reason to believe in such a distinct "metaphysical" modality. (There is a metaphysical modality, but it is just the logical modality.) There is nothing for a distinct metaphysical modality to explain; what is plausibly explainable is already explained. The only reason to believe in it is to justify certain antecedently held theoretical views (e.g. materialism); but this justification is circular in this context. The distinct modality is a frictionless wheel.

(viii) We do not even have a distinct concept of metaphysical modality for the second primitive to answer to. The momentary impression of such a concept stems from a false and confused understanding of such ontic/epistemic distinctions as that between apriority and necessity, and that between concept and property. But all that is easily subsumed under modal monism with the help of two-dimensional semantics.

(ix) Ultimately, there is just one circle of modal concepts, including both the rational modal concepts (validity, rational entailment, a priority, conceivability) and the metaphysical modal concepts (possibility, necessity, property). The relation between these is subtle, but it's a confusion to postulate a new primitive modality: all our reasons for believing in the metaphysical modality are grounded in the rational modality in the first place. (Witness the use of conceivability arguments in establishing a posteriori necessity.)

(10) LOOSE ENDS

To establish pure modal rationalism, we need to cross the bridge from a priority to necessity. This requires ruling out intermediate truths. If we can't rule out such truths, we have distinct circles of "positive" and "negative" rational modal concepts; so close yet so far! This isn't so crucial for the mind-body issue, but it's an important project in its own right.

3: THE CONTENT AND EPISTEMOLOGY OF PHENOMENAL BELIEF

[The material in this lecture now exists as a (much revised and expanded) full paper, ["The Content and Epistemology of Phenomenal Belief"](#).]

(1) INTRODUCTION

PHENOMENAL BELIEF = a belief about an experience (or about experiences in general). I'll concentrate on first-person phenomenal beliefs. E.g.

- "I am having a red experience now".
- "I am in pain".

PROJECT: Try to understand the content of such beliefs. And try to understand their first-person epistemology.

STARTING POINT: I assume "phenomenal realism". Roughly: that phenomenal properties (or qualia) exist - properties capturing what it is like to be in a given mental state - and that the phenomenal isn't conceptually reducible to the functional. On this view, when Mary learns what it is like to see something red, she learns something factual; and it's coherently conceivable that there be a functional (and environmental) duplicate of a conscious being with qualitatively different experiences.

[Includes: e.g. property dualism, qualia-friendly (type-B) materialism. Excludes e.g. eliminativism, analytic functionalism (type-A materialism).]

Q: Does this conceptual irreducibility of the phenomenal to the functional "spread" to other aspects of mentality, e.g. belief content?

A: (to be justified): Yes, at least for phenomenal beliefs.

(2) THE CONTENT OF PHENOMENAL CONCEPTS AND PHENOMENAL BELIEFS

I look at a tomato, have a red experience, and think: "I'm having an experience of such-and-such quality". There are various concepts of the quality in question that might yield a true belief.

THE COMMUNITY RELATIONAL CONCEPT: red_C = "the quality caused in normal observers in my community by red things" (i.e. by paradigms X, Y)

THE INDIVIDUAL RELATIONAL CONCEPT: red_I = "the quality normally caused in me by red things"

THE INDEXICAL CONCEPT: E = "the quality I'm experiencing now" (or: I'm ostending now)

THE QUALITATIVE CONCEPT: R = [red splotch]

The qualitative concept is the one that captures Mary's new knowledge: her new knowledge that certain experiences have such-and-such quality is knowledge that those experiences are R. This concept differs

from all the others: witness the *nontriviality* of:

- $\text{red_C} = R$
- $\text{red_I} = R$
- $E = R$

The knowledge Mary *gains* is that e.g.

- tomatoes look R to her (and probably to others)
- red things look R to her
- she is now having an R experience.

[Technically: red_C , red_I , E , and R , all have different primary intensions (across conceptually possible worlds), as witnessed by the aposteriority of these identities. They have the same secondary intensions.]

Contrast **Inverted Mary**, who has green experiences where Mary has red experiences. Her new knowledge differs from Mary's. She learns that tomatoes look G to her, that red things look G to her, and so on.

UPSHOT: If experience isn't conceptually supervenient on the functional, phenomenal belief isn't, either. (Mary and Inverted Mary are functional twins but have different beliefs.)

TEMPTATION: Handle the Mary/Inverted Mary difference analogously to standard externalism, e.g. the water/twin water case. (Oscar believes he's seeing water, Twin Oscar believes he's seeing twin water.) Roughly, similar "cognitive" or "notional" content ("the watery stuff around here"), different "relational content" (H_2O vs. XYZ).

BUT: This doesn't work. The "water" analogy works for red_C , red_I , or E . E.g. Mary and Twin Mary might both have concepts with the notional content of red_C but different relational content. But their R and G concepts differ in their notional content, not just their relational content.

Crucial factor: When Mary confidently believes that tomatoes look R , she can thereby rule out all epistemic possibilities (all centered worlds) in which tomatoes look G . When Oscar confidently believes that the glass contains water, he cannot thereby rule out epistemic possibilities in which the glass contains XYZ (unless he has some other knowledge, e.g. that $\text{water} = \text{H}_2\text{O}$). R is more analogous to a lucid concept of " H_2O " here, not the concept of "water". (Hence the disanalogy: one doesn't find twins with respective concepts of H_2O and XYZ .)

[Technically: Primary intension of "water" picks out the watery stuff near the center of a centered world. Primary intensions of red_C and red_I pick out the qualities satisfying the relevant descriptions near the center, and E picks out whatever quality is experienced or ostended by the center. But primary intension of R picks out (red splotch) in all worlds, and primary intension of G picks out (green splotch) in all worlds.]

In these cases: the quality of the experiences "gets inside" the notional content of the concept, and of the corresponding belief. The reference is somehow inside the sense, in a way much stronger than the usual "direct reference". I.e. the quality doesn't just determine secondary intension (as usual), but primary intension. (The only true "direct reference"? cf. Russell.) A strange and interesting phenomenon!

MORAL: Experience plays a role in *constituting* our phenomenal concepts and phenomenal beliefs. Mary's phenomenal concept/belief content supervenes not on her functioning but on her functioning plus phenomenology. Call this sort of concept an experience-constituted phenomenal concept. In the crucial cases (like Mary's), the phenomenal contribution to the content of a concept may be constituted by the quality of a single experience one is having right now. Call this a single-case experience-constituted phenomenal concept, or a **direct phenomenal concept** for short.

[Direct phenomenal concepts have the same primary intension and secondary intension, picking out experiences with the given quality in all worlds.]

A direct phenomenal concept need not accompany every experience, but they are common when one attends to one's experiences. One has an experience with a certain quality, and forms a concept of *this* sort of quality (N.B. not the mere indexical "this", as in E). The quality is "taken up" into the concept, such that the (primary intension of the) concept picks out experiences of that quality in any centered world.

A **direct phenomenal belief** arises when one forms a direct phenomenal concept based on a certain experience, and predicates that concept of the experience in question. E.g. one has a red experience, uses it to form the direct phenomenal concept R, and believes "this experience is R".

[Note: one can do all this in terms only of qualities, and not of "experiences" at all, if one doesn't like involving anything so object-like as an experience. One can e.g. invoke the direct phenomenal belief $E=E'$, where E is the indexical picking out the quality one is ostending now, and E' is a direct phenomenal concept constituted by that quality. But I do things in terms of experiences with qualities here, mostly because it makes for a more intuitive exposition.]

(3) THE EPISTEMOLOGY OF PHENOMENAL BELIEF

(3.1) LIMITED INCORRIGIBILITY THESIS

Direct phenomenal beliefs are *incorrigible*; i.e., they cannot be false. If a direct phenomenal concept R is constituted by a quality Q of experience E, R applies truly to any experience with Q. In a direct phenomenal belief, R is predicated of E, which has Q, so the belief is true. The role of the experience in constituting both concept and object ensures this. [cf. Pollock.]

Note 1: **LIMITATIONS:** Most phenomenal beliefs are still corrigible, as most phenomenal beliefs are not direct phenomenal beliefs. E.g. applications of a *pre-existing* phenomenal concept to a new situation ("I am having a red experience now", "I am in pain", etc.) are corrigible, as are applications of a direct phenomenal concept to an experience other than the one that constitutes it. All the standard

counterexamples fall into classes like these, so none are counterexamples to the limited incorrigibility thesis here. But I suggest that this thesis captures the *plausible core* of standard incorrigibility theses.

Other limitations: (a) Direct phenomenal beliefs are incorrigible, but subjects may be corrigible (at least in some circumstances) about whether they are having a direct phenomenal belief. (b) Most (token) experiences do not have corresponding direct phenomenal concepts, so are not incorrigibly known. (c) It may be that some experiences (e.g. fleeting or background experiences) *cannot* be taken up into a direct phenomenal concept, so are not incorrigibly knowable.

Note 2: **NONTRIVIALITY**: It might be thought that this incorrigible knowledge is trivial, like "this is this", or "I am here". But this is not so. The belief in question is *cognitively significant*: it heavily constrains the subject's space of epistemic possibilities, unlike the trivial beliefs in question. Technically: the belief in question has a conceptually contingent primary intension (false in many centered worlds), whereas the trivial beliefs have a conceptually necessary primary intension (true in all centered worlds, omitting a wrinkle or two).

(3.2) EPISTEMOLOGICAL PROBLEMS FOR PHENOMENAL REALISM

Phenomenal realism, especially property dualism, is often thought to face epistemological problems centering on the relation between experience and beliefs about experience. E.g. Shoemaker, "Functionalism and qualia", and "The paradox of phenomenal judgment", in *The Conscious Mind*, Chapter 5.

Starkest argument (for the epiphenomenalist): if experience is causally irrelevant, then experiences are not causally related to beliefs about experiences, so beliefs about experiences are not knowledge. **Reply**: The relation between experience and phenomenal belief is *tighter* than any causal relation; it is constitution. And it is the constitutive relation that qualifies a phenomenal belief as knowledge.

Related argument: A zombie would have the same beliefs as me, caused by the same mechanisms, but its beliefs are not justified, so my beliefs are not justified. **Reply**: The premise is false. My zombie twin's relevant beliefs are not the same as mine, because of the constitutive role of experience in phenomenal belief. (His "direct" phenomenal concepts are arguably empty, or close to empty, if he has concepts at all.)

[In *The Conscious Mind* I handled these problems differently, relying mostly on our "acquaintance" with experience, and relegating the constitutive relation to a minor role. I now think the constitutive relation is central in answering the arguments. Acquaintance still has an important role to play, but this is a further story.]

(3.3) "THE MYTH OF THE GIVEN"

Presumably this view accepts some sort of "given". *Question*: Are the arguments that have been put

forward against the "myth of the given" good arguments against this view?

(I) Sellars' inconsistent triad.

- (A) "S senses red sense-content s" entails "S non-inferentially knows that s is red";
- (B) The ability to sense sense-contents is unacquired.
- (C) The ability to know facts of the form "x is phi" is acquired.

My view denies (A). An experience need not be accompanied by a phenomenal belief, and most experiences are not. Phenomenal beliefs are more cognitively sophisticated than experiences. The belief involves concepts; the experience does not, or need not. Knowledge requires belief, so experience does not entail phenomenal knowledge.

Sellars notes that denying (A) entails that a sensing doesn't constitute an item of knowledge. I think this is correct (although it can *partly* constitute an item of knowledge). Sellars associates the "given" most strongly with (A), so doesn't explicitly argue further against (A)-denying views here.

But such a view leads into

(II) The justification dilemma. If experience is nonconceptual, how can it play a role in justifying something conceptual? (Davidson, Bonjour, McDowell.)

- (1) There can be no **inferential** relation between an experience and a belief, as inference requires connections within the conceptual domain.
- (2) But a mere **causal** relation between experience and belief cannot justify the belief. (McDowell: "This offers exculpation where we wanted justification.")
- (3) So: experience cannot play any foundational role in the justification of belief?

REPLY: An option has been missed. The relation between experience and belief is not inference, not causation, not identity, but (partial) constitution. It is the role an experience plays in constituting a direct phenomenal belief that makes that belief incorrigible, and indeed incorrigible in virtue of its phenomenological structure, and so justified.

In any case, the view avoids Sellars' central version of the "given" (i.e., (A)) and McDowell's central version (i.e. a mere causal relation), and the arguments against those versions. Are there further arguments?

FURTHER QUESTIONS:

(1) Does the experience-dependence and functional irreducibility of phenomenal concepts and beliefs extend to other concepts and beliefs?

[Suggestive, at least for perceptual concepts and those based on them.]

(2) Does the limited foundationalism for direct phenomenal beliefs support a wider foundationalism, for a wider class of beliefs?

[At least the gap between experience and belief - often taken to be the greatest problem for foundationalism - has been bridged. The gap between direct phenomenal beliefs and other phenomenal beliefs is bridgeable, by invoking "standing phenomenal concepts" (the sort of coarse-grained qualitative concept we can have of a quality when that quality isn't present). The other major gap, between phenomenal beliefs and beliefs about the external world, still remains.]

The Content and Epistemology of Phenomenal Belief

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

1 Introduction

Experiences and beliefs are different sorts of mental states, and are often taken to belong to very different domains. Experiences are paradigmatically phenomenal, characterized by what it is like to have them. Beliefs are paradigmatically intentional, characterized by their propositional content. But there are a number of crucial points where these domains intersect. One central locus of intersection arises from the existence of phenomenal beliefs: beliefs that are about experiences.

The most important phenomenal beliefs are *first-person* phenomenal beliefs: a subject's beliefs about his or her own experiences, and especially, about the phenomenal character of the experiences that he or she is currently having. Examples include the belief that one is now having a red experience, or that one is experiencing pain.

These phenomenal beliefs raise important issues in metaphysics, in the theory of content, and in epistemology. In metaphysics, the relationship between phenomenal states and phenomenal beliefs is sometimes taken to put strong constraints on the metaphysics of mind. In the theory of content, analyzing the content of phenomenal beliefs raises special issues for a general theory of content to handle, and the content of such beliefs has sometimes been taken to be at the foundations of a theory of content more generally. In epistemology, phenomenal beliefs are often taken to have a special epistemic status, and are sometimes taken to be the central epistemic nexus between cognition and the external world.

My project here is to analyze phenomenal beliefs in a way that sheds some light on these issues. I will start by focusing on the content of these beliefs, and will use the analysis developed there to discuss the underlying factors in virtue of which this content is constituted. I will then apply this framework to the central epistemological issues in the vicinity: incorrigibility, justification, and the dialectic over the "Myth of the Given".

1.1 Phenomenal realism

The discussion that follows is premised upon what I call "phenomenal realism": the view that there are phenomenal properties (or phenomenal qualities, or qualia), properties that type mental states by what it is like to have them, and that phenomenal properties are not conceptually reducible to physical or functional properties (or equivalently, that phenomenal concepts are not reducible to physical or functional concepts). On this view, there are truths about what it is like to be a subject that are not entailed a priori by the physical and functional truth (including the environmental truth) about that subject.

The phenomenal realist view is most easily illustrated with some familiar thought-experiments. Consider Frank Jackson's case of Mary, the neuroscientist who knows all relevant physical truths about color processing, but whose visual experience has been entirely monochromatic. On the phenomenal realist view, Mary lacks factual knowledge concerning what it is like to see red. Views that deny this deny phenomenal realism. Or consider cases in which a hypothetical being has the same physical, functional and environmental properties as an existing conscious being, but does not have the same phenomenal properties. Such a being might be a zombie, lacking experiences altogether, or it might be an inverted being, with experiences of a different character. On the phenomenal realist view, some such duplicates are coherently conceivable, in the sense that there is no a priori contradiction in the hypothesis in question. Views that deny this deny phenomenal realism.

(What if someone holds that functional duplicates without consciousness are coherently conceivable, but that physical duplicates without consciousness are not? Such a view would be in the spirit of phenomenal realism. This suggests that we could define phenomenal realism more weakly as the thesis that the phenomenal is not conceptually reducible to the functional, omitting mention of the physical. I do not define it this way, for two reasons. First, I think if functional duplicates without consciousness are conceivable, physical duplicates without consciousness must be conceivable too, as there is no reasonable possibility of a conceptual entailment from microphysical to phenomenal that does not proceed via the functional. Second, it is not easy to give a precise account of what functional duplication consists in, and stipulating physical identity finesses that question. But if someone disagrees, everything that I say in this paper will apply, with appropriate changes, on the weaker view.)

Phenomenal realism subsumes most varieties of dualism about the phenomenal. It also subsumes many varieties of materialism. In particular it subsumes what I have called "type-B" materialism (see Chalmers 2002a): views that hold that there is an a posteriori necessary entailment from the physical to the phenomenal, so that there is an epistemic or conceptual gap between the physical and phenomenal domains, but no ontological gap. Views of this sort typically allow that Mary gains factual knowledge when she sees red for the first time, but hold that it is knowledge of an old fact known in a new way; and they typically hold that the duplication cases mentioned above are conceptually coherent but not metaphysically possible.

Phenomenal realism excludes what I have called "type-A" materialism: views that hold that all phenomenal truths are entailed a priori by physical truths. Such views include eliminativism about the phenomenal, as well as analytical functionalism and logical behaviorism, and certain forms of analytic

representationalism. Views of this sort typically deny that Mary gains any knowledge when she sees red for the first time, or hold that she gains only new abilities; and they typically deny that the duplication cases mentioned above are coherently conceivable.

Those who are not phenomenal realists might want to stop reading now, but there are two reasons why they might continue. First, although the arguments I will give for my view of phenomenal beliefs will presuppose phenomenal realism, it is possible that some aspects of the view itself may be tenable even on some views that deny phenomenal realism. Second, some of the most important arguments *against* phenomenal realism are epistemological arguments that center on the connection between experience and belief. I will be using my analysis to help rebut those arguments, and thus indirectly to support phenomenal realism against its opponents.

A note on modality: Because I am assuming phenomenal realism but not property dualism in this paper, all references to necessity and possibility should be taken as invoking conceptual necessity and possibility. Similarly, talk of possible worlds can be taken as invoking conceivable worlds (corresponding to the epistemically constructed scenarios of Chalmers (forthcoming); see also the appendix), and talk of constitutive relations should be taken as invoking conceptually necessary connections. If one accepts a certain sort of link between conceptual and metaphysical possibility (e.g. the thesis that ideal primary conceivability entails primary possibility), then these references can equally be taken as invoking metaphysical possibility and necessity.

A note on phenomenal properties: It is natural to speak as if phenomenal properties are instantiated by mental states, and as if there are entities, experiences, that bear their phenomenal properties essentially. But one can also speak as if phenomenal properties are directly instantiated by conscious subjects, typing subjects by aspects of what it is like to be them at the time of instantiation. These ways of speaking do not commit one to corresponding ontologies, but they at least suggest such ontologies. In a *quality-based* ontology, the subject-property relation is fundamental. From this one can derive a subject-experience-property structure, by identifying experiences with phenomenal states (instantiations of phenomenal properties), and attributing phenomenal properties to these states in a derivative sense. In a more complex *experience-based* ontology, a subject-experience-property structure is fundamental (where experiences are phenomenal individuals, or at least something more than property instantiations), and the subject-property relation is derivative. In what follows, I will sometimes use both sorts of language, and will be neutral between the ontological frameworks.

2 The Content of Phenomenal Concepts and Phenomenal Beliefs

2.1 Relational, demonstrative, and pure phenomenal concepts

Phenomenal beliefs involve the attribution of phenomenal properties. These properties are attributed under phenomenal concepts. To understand the content of phenomenal beliefs, we need to understand the nature and content of phenomenal concepts.

I look at a red apple, and visually experience its color. This experience instantiates a phenomenal quality R, which we might call phenomenal redness. It is natural to say that I am having a red experience, even though of course experiences are not red in the same sense in which apples are red. Phenomenal redness (a property of experiences, or of subjects of experience) is a different property from external redness (a property of external objects), but both are respectable properties in their own right.

I attend to my visual experience, and think *I am having an experience of such-and-such quality*, referring to the quality of phenomenal redness. There are various concepts of the quality in question that might yield a true belief.[*]

*[[I take concepts to be mental entities on a par with beliefs: they are constituents of beliefs (and other propositional attitudes) in a manner loosely analogous to the way in which words are constituents of sentences. Like beliefs, concepts are tokens rather than types in the first instance. But they also fall under types, some of which I explore in what follows. In such cases it is natural to use singular expressions such as 'the concept' for a concept-type, just as one sometimes uses expressions such as 'the belief' for a belief-type, or 'the word' for a word-type.]]

We can first consider the concept expressed by 'red' in the public-language expression 'red experience', or the concept expressed by the public-language expression 'phenomenal redness'. The reference of these expressions is fixed via a relation to red things in the external world, and ultimately via a relation to certain paradigmatic red objects that are ostended in learning the public-language term 'red'. A language learner learns to call the experiences typically brought about by these objects 'red' (in the phenomenal sense), and to call the objects that typically bring about those experiences 'red' (in the external sense). So the phenomenal concept involved here is *relational*, in that it has its reference fixed by a relation to external objects. The property that is referred to need not be relational, however. The phenomenal concept plausibly designates an intrinsic property rigidly, so that there are counterfactual worlds in which red experiences are never caused by red things.

One can distinguish at least two relational phenomenal concepts, depending on whether reference is fixed by relations across a whole community of subjects, or by relations restricted to the subject in question. The first is what we can call the *community relational concept*, or *red_C*. This can be glossed roughly as *the phenomenal quality typically caused in normal subjects within my community by paradigmatic red things*. The second is what we can call the *individual relational concept*, or *red_I*. This can be glossed roughly as *the phenomenal quality typically caused in me by paradigmatic red things*. The two concepts *red_C* and *red_I* will corefer for normal subjects, but for abnormal subjects they may yield different results. For example, a red/green-inverted subject's concept *red_C* will refer to (what others call) phenomenal redness, but his or her concept *red_I* will refer to (what others call) phenomenal greenness.

The public-language term 'red' as a predicate of experiences can arguably be read as expressing either *red_C* or *red_I*. The community reading of 'red' guarantees a sort of shared meaning within the community, in that all uses of the term are guaranteed to corefer, and in that tokens of sentences such as 'X has a red experience at time t' will have the same truth-value whether uttered by normal or abnormal subjects. On the other hand, the individual reading allows a subject better access to the term's referent. On this reading, an unknowingly inverted subject's term 'red' will refer to what she think it refers to

(unless the inversion was recent), while on the community reading, her term 'red' may refer to something quite different, and her utterance 'I have had red experiences' may even be unknowingly quite false.[*] In any case, we need not settle here just what is expressed by phenomenal predicates in public language. All that matters is that both concepts are available.

*[[These cases may not be entirely hypothetical. Nida-Rumelin (1996) gives reasons, based on the neurobiological and genetic bases of colorblindness, to believe that a small fraction of the population may actually be spectrum-inverted with respect to the rest of us. If so, it is natural to wonder just what their phenomenal expressions refer to.]]

Phenomenal properties can also be picked out by demonstratives. When seeing the tomato, I can refer demonstratively to a visual quality associated with it, using a concept I might express by saying 'this quality' or 'this sort of experience'. The relevant concept here is the phenomenal demonstrative that we might call *E*, which functions roughly by picking out whatever quality the subject is currently ostending. Like other demonstratives, it has a "character", which fixes reference in a context roughly by picking out whatever quality is ostended in that context; and it has a distinct "content", corresponding to the quality that is actually ostended - in this case, phenomenal redness. The demonstrative concept *E* rigidly designates its referent, so that it picks out the quality in question even in counterfactual worlds in which no-one is ostending the quality.

The three concepts *red_C*, *red_I*, and *E* may all refer to the same quality, phenomenal redness. All of them fix reference to phenomenal redness relationally, characterizing it in terms of its relations to external objects or acts of ostension. There is another crucial phenomenal concept in the vicinity, one that does not pick out phenomenal redness relationally, but rather picks it out directly, in terms of its intrinsic phenomenal nature. This is what we might call a *pure phenomenal concept*.

To see the need for the pure phenomenal concept, consider the knowledge that Mary gains when she learns for the first time what it is like to see red. She learns that seeing red has such-and-such quality. Mary learns (or reasonably comes to believe) that red things will typically cause experiences of such-and-such quality in her, and in other members of her community. She learns (or gains the cognitively significant belief) that the experience she is now having has such-and-such quality, and that the quality she is now ostending is such-and-such. Call Mary's "such-and-such" concept here *R*.

Mary's concept *R* picks out phenomenal redness, but it is quite distinct from the concepts *red_C*, *red_I*, and *E*. We can see this by using cognitive significance as a test for difference between concepts. Mary gains the belief *red_C*=*R* - that the quality typically caused in her community by red things is such-and-such - and this belief is cognitively significant knowledge. She gains the cognitively significant belief *red_I* = *R* in a similar way. And she gains the belief *E*=*R* - roughly, that the quality she is now ostending is such-and-such.

Mary's belief *E*=*R* is as cognitively significant as any other belief in which the object of a demonstrative is independently characterized: e.g. my belief *I am David Chalmers*, or my belief *that object is tall*, or my belief *that shape is roundness*. For Mary, *E*=*R* is not a priori. No a priori reasoning can rule out the hypothesis that she is now ostending some other quality entirely, just as no a priori reasoning can rule out

the hypothesis that I am David Hume, or that the object I am pointing to is short. Indeed, nothing known a priori entails that R is ever instantiated in the actual world.

So the concept R is quite distinct from red_C , red_I , and E . Unlike the other concepts, the pure phenomenal concept characterizes the phenomenal quality *as* the phenomenal quality that it is.

The concept R is difficult to express directly in language, since the most natural terms, such as 'phenomenal redness' and 'this experience', arguably express other concepts such as red_C and E . Still, one can arguably discern uses of these terms that express pure phenomenal concepts; or if not, one can stipulate such uses. For example, Chisholm suggests that there is a "noncomparative" sense of expressions such as 'looks red'; this sense seems to express a pure phenomenal concept, whereas his "comparative sense" seems to express a relational phenomenal concept.[*] And at least informally, demonstratives are sometimes used to express pure phenomenal concepts. For example, the belief that $E=R$ might be informally expressed by saying something like "this quality is *this* quality".

*[[The distinction also roughly tracks Nida-Rumelin's (1995; 1997) distinction between "phenomenal" and "nonphenomenal" readings of belief attributions concerning phenomenal states. "Phenomenal" belief attributions seem to require that the subject satisfies the attribution by virtue of a belief involving a pure phenomenal concept, while "nonphenomenal" attributions allow that the subject can satisfy the attribution by virtue of a belief involving a relational phenomenal concept.]]

It may be that there is a sense in which R can be regarded as a "demonstrative" concept. I will not regard it this way: I take it that demonstrative concepts work roughly as analyzed by Kaplan (1977), so that they have an epistemic reference-fixing "character" that leaves their referent open. This is how E behaves: its content might be expressed roughly as "this quality, whatever it happens to be". R , on the other hand, is a substantive concept that is tied *a priori* to a specific sort of quality, so it does not behave the way that Kaplan suggests that a demonstrative should. Still, there is an intimate relationship between pure and demonstrative phenomenal concepts that I will discuss later in the paper; and if someone wants to count pure phenomenal concepts as "demonstrative" in a broad sense, there is no great harm in doing so, as long as the relevant distinctions are kept clear. What matters for my purposes is not the terminological point, but the more basic point that the distinct concepts E and R exist. For a phenomenal realist, there can be no doubt about this.[*]

The relations among these concepts can be analyzed straightforwardly using the two-dimensional framework for representing the content of concepts. A quick introduction to this framework is given in an appendix; more details can be found in Chalmers (2002c). The central points of this paper should be comprehensible if matters involving the two-dimensional framework are skipped, but the framework makes the analysis of some crucial points much clearer.

According to the two-dimensional framework, when an identity $A=B$ is a posteriori, the concepts A and B have different epistemic (or primary) intensions. If A and B are rigid concepts and the identity is true, A and B have the same subjunctive (or secondary) intensions. So we should expect that the concepts red_C , red_I , E , and R have different epistemic intensions, but the same subjunctive intension. And this is what

we find. The subjunctive intension of each picks out phenomenal redness in all worlds. The epistemic intension of *red_C* picks out, in a given centered world, roughly the quality typically caused by certain paradigmatic objects in the community of the subject at the center of the world. The epistemic intension of *red_I* picks out roughly the quality typically caused by those objects in the subject at the center.

As for the demonstrative concept *E*: to a first approximation, one might hold that its epistemic intension picks out the quality that is ostended by the subject at the center. This characterization is good enough for most of our purposes, but it is not quite correct. It is possible to ostend two experiences simultaneously and invoke two distinct demonstrative concepts, as when one thinks *that quality differs from that quality*, ostending two different parts of a symmetrical visual field (see Austin 1990). Here no descriptive characterization such as the one above will capture the difference between the two concepts. It is better to see *E* as a sort of indexical, like *I* or *now*. To characterize the epistemic possibilities relevant to demonstrative phenomenal concepts, we need centered worlds whose centers contain not only a "marked" subject and time, but also one or more marked experiences; in the general case, a sequence of such experiences.[*] Then a concept such as *E* will map a centered world to the quality of the "marked" experience (if any) in that world. Where two demonstrative concepts *E1* and *E2* are involved, as above, the relevant epistemic possibilities will contain at least two marked experiences, and we can see *E1* as picking out the quality of the first marked experience in a centered world, and *E2* as picking out the quality of the second. Then the belief above will endorse all worlds at which the quality of the first marked experience differs from the quality of the second. This subtlety will not be central for the purposes of this paper.

*[[In the experience-based framework: if experiences do not map one-to-one to instances of phenomenal properties, then instances of phenomenal properties should be marked instead.]]

The epistemic intension of *R* is quite distinct from all of these. It picks out phenomenal redness in all worlds. I will analyze this matter in more depth shortly; but one can see intuitively why this is plausible. When Mary believes *roses cause R experiences*, or *I am currently having an R experience*, she thereby excludes all epistemic possibilities in which roses cause some other quality (such as *G*, phenomenal greenness), or in which she is experiencing some other quality: only epistemic possibilities involving phenomenal redness remain.

The cognitive significance of identities such as *red_C=R*, *red_I=R*, and *E=R* is reflected in the differences between the concept's epistemic intensions. The first two identities endorse all epistemic possibilities in which paradigmatic objects stand in the right relation to experiences of *R*; these are only a subset of the epistemic possibilities available a priori. The third identity endorses all epistemic possibilities in which the marked experience at the center (or the ostended experience, on the rough characterization) is *R*. Again, there are many epistemic possibilities (a priori) that are not like this: centered worlds in which the marked experience is *G*, for example. Once again, this epistemic contingency reflects the cognitive significance of the identity.[*]

(Phenomenal realists (e.g. Loar 1997, Hawthorne 2001) analyzing what Mary learns have occasionally suggested that her phenomenal concept is a demonstrative concept. This is particularly popular as a way

of resisting anti-materialist arguments, as it is tempting to invoke the distinctive epistemic and referential behavior of demonstrative concepts in explaining why an epistemic gap does not reflect an ontological gap. But on a closer look it is clear that Mary's central phenomenal concept R (the one that captures what she learns) is *distinct* from her central demonstrative concept E , as witnessed by the nontrivial identity $E=R$, and is not a demonstrative concept in the usual sense. This is not just a terminological point. Those who use these analyses to rebut anti-materialist arguments typically rely on analogies with the epistemic and referential behavior of ordinary (Kaplan-style) demonstratives. Insofar as these analyses rely on such analogies, they fail. Something similar applies to analyses that liken phenomenal concepts to indexical concepts (e.g. Ismael 1999, Perry 2001). If my analysis is correct, then pure phenomenal concepts (unlike demonstrative phenomenal concepts) are not indexical concepts at all.)

2.2 Inverted Mary

We can now complicate the situation by introducing another thought-experiment on top of the first one. Consider the case of *Inverted Mary*, who is physically, functionally, and environmentally just like Mary, except that her phenomenal color vision is red/green inverted. (I will assume for simplicity that Inverted Mary lives in a community of inverted observers.) Like Mary, Inverted Mary learns something new when she sees red things for the first time. But Inverted Mary learns something different from what Mary learns. Where Mary learns that tomatoes cause experiences of (what we call) phenomenal redness, Inverted Mary learns that they cause experiences of (what we call) phenomenal greenness. In the terms given earlier, Mary acquires beliefs $red_C=R$, $red_I=R$, and $E=R$, while Inverted Mary acquires beliefs $red_C=G$, $red_I=G$, and $E=G$ (where G is the obvious analog of R). So Mary and Inverted Mary acquires beliefs with quite different contents.

This is already enough to draw a strong conclusion about the irreducibility of content. Recall that Mary and Inverted Mary are physical/functional and environmental twins, even after they see red things for the first time. Nevertheless, they have beliefs with different contents. It follows that belief content does not supervene conceptually on physical/functional properties. And it follows from this that intentional properties are not conceptually supervenient on physical/functional properties, in the general case.

This is a nontrivial conclusion. Phenomenal realists often hold that while the phenomenal is conceptually irreducible to the physical and functional, the intentional can be analyzed in functional terms. But if what I have said here is correct, then this irreducibility cannot be quarantined in this way. If the phenomenal is conceptually irreducible to the physical and functional, so too is at least one aspect of the intentional: the content of phenomenal beliefs.

At this point, there is a natural temptation to downplay this phenomenon by reducing it to a sort of dependence of belief content on reference that is found in many other cases: in particular in the cases that are central to externalism about the content of belief. Take Putnam's case of Twin Earth. Oscar and Twin Oscar are functional duplicates, but they inhabit different environments: Oscar's contains H₂O as the clear liquid in the oceans and lakes, while Twin Oscar's contains XYZ (which we count not as water but as twin water). As a consequence, Oscar's *water* concept refers to water (H₂O), while Twin Oscar's

analogous concept refers to twin water (XYZ). Because of this difference in reference, Oscar and Twin Oscar seem to have different beliefs: Oscar believes that water is wet, while Twin Oscar believes that twin water is wet. Perhaps the case of Mary and Inverted Mary is just like this?[*]

*[[This sort of treatment of phenomenal belief is suggested by Francescotti (1994).]]

The analogy does not go through, however. Or rather, it goes through only to a limited extent. Oscar and Twin Oscar's *water* concepts here are analogous to Mary and Inverted Mary's relational phenomenal concepts (*red_C* or *red_I*), or perhaps to their demonstrative concepts. For example, the relational concepts that they express with their public-language expressions 'red experience' will refer to two different properties, phenomenal redness and phenomenal greenness. Mary and Inverted Mary can deploy these concepts in certain beliefs, such as the beliefs that they express by saying 'Tomatoes cause red experiences', even before they leave their monochromatic rooms for the first time. Because of the distinct referents of their concepts, there is a natural sense (Nida-Rumelin's "nonphenomenal" sense) in which we can say that Mary believed that tomatoes caused red experiences, while Inverted Mary did not; she believed that tomatoes caused green experiences. Here the analogy goes through straightforwardly.

The pure phenomenal concepts *R* and *G*, however, are less analogous to the two *water* concepts than to the chemical concepts *H2O* and *XYZ*. When Oscar learns the true nature of water, he acquires the new belief *water* = *H2O*, while Twin Oscar acquires an analogous belief involving *XYZ*. When Mary learns the true nature of red experiences, she acquires a new belief *red_C* = *R*, while Inverted Mary acquires an analogous belief involving *G*. That is, Mary and Inverted Mary's later knowledge involving *R* and *G* is fully lucid knowledge of the referents of the concepts in question, analogous to Oscar and Twin Oscar's knowledge involving the chemical concepts *H2O* and *XYZ*.

But here we see the strong disanalogy. Once Oscar acquires the chemical concept *H2O* and Twin Oscar acquires *XYZ*, they will no longer be twins: their functional properties will differ significantly. By contrast, at the corresponding point Mary and Inverted Mary are still twins. Even though Mary has the pure phenomenal concept *R* and Inverted Mary has *G*, their functional properties are just the same. So the difference between the concepts *R* and *G* across functional twins is something that has no counterpart in the standard Twin Earth story.

All this reflects the fact that in standard externalist cases, the pairs of corresponding concepts may differ in reference, but they have the same or similar *epistemic* or *notional* contents. Oscar and Twin Oscar's *water* concepts have different referents (*H2O* vs. *XYZ*), but they have the same epistemic contents: both intend to refer to roughly the liquid around them with certain superficial properties. Something like this applies to Mary's and Inverted Mary's relational phenomenal concepts, which have different referents but the same epistemic content (which picks out whatever quality stands in a certain relation), and to their demonstrative concepts (which pick out roughly whatever quality happens to be ostended).

In terms of the two-dimensional framework, where epistemic contents correspond to epistemic intensions: Oscar's and Twin Oscar's *water* concepts have the same epistemic intension but different

subjunctive intensions. A similar pattern holds in all the cases characteristic of standard externalism. The pattern also holds for Mary's and Inverted Mary's relational phenomenal concepts, and their demonstrative phenomenal concepts.

But Mary's concept *R* and Twin Mary's concept *G* have *different* epistemic contents. In this way they are analogous to Oscar's concept *H₂O* and Twin Oscar's concept *XYZ*. But again, the disanalogy is that *R* and *G* are possessed by twins, and *H₂O* and *XYZ* are not. So the case of Inverted Mary yields an entirely different phenomenon: a case in which *epistemic* content differs between twins.

This can be illustrated by seeing how the concepts in question are used to constrain epistemic possibilities. When Oscar confidently believes that there is water in the glass, he is not thereby in a position to rule out the epistemic possibility that there is *XYZ* in the glass (unless he has some further knowledge, such as the knowledge that water is *H₂O*). The same goes for Twin Oscar's corresponding belief. For both of them, it is equally epistemically possible that the glass contains *H₂O* and that it contains *XYZ*. Any epistemic possibility compatible with Oscar's belief is also compatible with Twin Oscar's belief: in both cases, these will be roughly those epistemic possibilities in which a sample of the dominant watery stuff in the environment is in the glass.

Epistemic content reflects the way that a belief constrains the space of epistemic possibilities, so Oscar's and Twin Oscar's epistemic contents are the same. Something similar applies to Mary and Inverted Mary, at least where their pairwise relational and demonstrative phenomenal concepts are concerned. When Mary confidently believes (under her relational concept) that her mother is having a red experience, for example, she is not thereby in a position to rule out the epistemic possibility that her mother is having an experience with the quality *G*. Both Mary's and Inverted Mary's beliefs are compatible with any epistemic possibility in which the subject's mother is having the sort of experience typically caused in the community by paradigmatic red objects. So their beliefs have the same epistemic contents.

But Mary's and Inverted Mary's pure phenomenal concepts do not work like this. Mary's concept *R* and Inverted Mary's concept *G* differ not just in their referents but in their epistemic contents. When Mary leaves the monochromatic room and acquires the confident belief (under her pure phenomenal concept) that tomatoes cause red experiences, she is thereby in a position to rule out the epistemic possibility that tomatoes cause experiences with quality *G*. The only epistemic possibilities compatible with her belief are those in which tomatoes cause *R* experiences. For Inverted Mary, things are reversed: the only epistemic possibilities compatible with her belief are those in which tomatoes cause *G* experiences. So their epistemic contents are quite different.

Again, the epistemic situation with *R* and *G* is analogous to the epistemic situation with the concepts *H₂O* and *XYZ*. When Oscar believes (under a fully lucid chemical concept) that the glass contains *H₂O*, he is thereby in a position to rule out all epistemic possibilities in which the glass contains *XYZ*. For Twin Oscar, things are reversed. This is to say that *H₂O* and *XYZ* have different epistemic contents. The same goes for *R* and *G*.

So in the case of the pure phenomenal concepts, uniquely, we have a situation in which two concepts differ in their epistemic content despite the subjects being physically identical. So phenomenal concepts seem to give a case in which even epistemic content is not conceptually supervenient on the physical.

Using the two-dimensional framework: the epistemic intension of a concept reflects the way it applies to epistemic possibilities. We saw above that the epistemic intensions of Oscar's and Twin Oscar's *water* concepts are the same, as are the epistemic intensions of Mary's and Inverted Mary's relational and demonstrative phenomenal concepts. But *R* and *G* differ in the way they apply to epistemic possibilities, and their epistemic intensions differ accordingly: the epistemic intension of *R* picks out phenomenal redness in all worlds, and the epistemic intension of *G* picks out phenomenal greenness in all worlds. When Mary thinks *I am having an R experience now*, the epistemic intension of her thought is true at all and only those worlds in which the being at the center is having an R experience.

Something very unusual is going on here. In standard externalism, and in standard cases of so-called "direct reference", a referent plays a role in constituting the subjunctive content (subjunctive intension) of concepts and beliefs, while leaving the epistemic content (epistemic intension) unaffected. In the pure phenomenal case, by contrast, the quality of the experiences plays a role in constituting the *epistemic* content of the concept and of the corresponding belief. One might say very loosely that in this case, the referent of the concept is somehow present inside the concept's sense, in a way much stronger than in the usual cases of "direct reference".[*]

*[[Martine Nida-Rumelin suggests in a forthcoming paper that we can call this sort of concept a *super-rigid* concept. A rigid concept is one whose subjunctive intension picks out the same referent at all worlds; a super-rigid concept is a rigid concept whose epistemic intension picks out the same referent at all worlds.]]

One might see here some justification for Russell's claim that we have a special capacity for direct reference to our experiences.[*] Contemporary direct reference theorists hold that Russell's view was too restrictive, and that we can make direct reference to a much broader class of entities. But the cases they invoke are "direct" only in the weak sense outlined above: the subjunctive content depends on the referent, but the epistemic content of the concept does not. In the phenomenal case, the epistemic content itself seems to be constituted by the referent. It is not hard to imagine that some such epistemic requirement on direct reference is what Russell had in mind.

*[[Russell also held that direct reference is possible to universals, and perhaps to the self. It is arguable that for at least some universals (in the domains of causation, space, time, or mathematics, perhaps), one can form a concept whose epistemic content picks out instances of that universal in all worlds. So there is at least a limited analogy here, though it seems unlikely that the content of such a (token) concept is directly constituted by an underlying instance of the universal, in the manner suggested below.

There is no analogous phenomenon with the self. There may, however, be a different sense in which we can make "direct reference" to the self, to the current time, and to particular experiences: this is the sort of direct indexical reference that corresponds to the need to build these entities into the center of a centered world. We can refer to these "directly" (in a certain sense) under indexical concepts; but we cannot form concepts whose epistemic content reflect the referents in question. This suggests that direct reference to particulars and direct reference to properties are quite different

phenomena.]]

3 The Constitution of Phenomenal Beliefs

3.1 Direct phenomenal concepts and beliefs

We have seen that the content of phenomenal concepts and phenomenal beliefs does not supervene conceptually on physical properties. Does this content supervene conceptually on some broader class of properties, and if so, on which? I will offer an analysis of how the content of pure phenomenal concepts is constituted. I will not give a knockdown argument for this analysis by decisively refuting all alternatives, but I will offer it as perhaps the most natural and elegant account of the phenomena, and as an account which can in turn do further explanatory work.

To start with, it is natural to hold that the content of phenomenal concepts and beliefs supervenes conceptually on the combination of physical and phenomenal properties. Mary and Inverted Mary are physical twins, but they are phenomenally distinct, and this phenomenal distinctness (Mary experiences phenomenal redness, Inverted Mary experiences phenomenal greenness) precisely mirrors their intentional distinctness (Mary believes that tomatoes cause R experiences, Inverted Mary believes that tomatoes cause G experiences). It is very plausible to suppose that their intentional distinctness holds in virtue of their phenomenal distinctness.

The alternative is that the intentional content of the phenomenal concept is conceptually independent of both physical and phenomenal properties. If that is so, it should be conceivable that two subjects have the same physical and phenomenal properties, while having phenomenal beliefs that differ in content. Such a case might involve Mary and Mary' as physical and phenomenal twins, who are both experiencing phenomenal redness for the first time (while being phenomenally identical in all other respects), with Mary acquiring the belief that tomatoes cause R experiences while Mary' acquires the belief that tomatoes cause G experiences. It is not at all clear that such a case is conceivable.

Another possibility is that the intentional content of Mary's phenomenal concept in question might be determined by phenomenal states *other* than the phenomenal redness that Mary is visually experiencing. For example, maybe Mary's belief content is determined by a faint phenomenal "idea" that goes along with her phenomenal "impression", where the former is not conceptually determined by the latter, and neither is conceptually determined by the physical. In that case, it should once again be conceivable that twins Mary and Mary' both visually experience phenomenal redness upon leaving the room, with Mary acquiring the belief that tomatoes cause R experiences while Mary' acquires the belief that tomatoes cause G experiences, this time because of a difference in their associated phenomenal ideas. But again, it is far from clear that this is conceivable.

There is a very strong intuition that the content of Mary's phenomenal concept and phenomenal belief is *determined* by the phenomenal character of her visual experience, in that it will vary directly as a function of that character in cases where that character varies while physical and other phenomenal

properties are held fixed, and that it will not vary independently of that character in such cases. I will adopt this claim as a plausible working hypothesis.

In particular, I will take it that in cases like Mary's, the content of a phenomenal concept and a corresponding phenomenal belief, is partly *constituted* by an underlying phenomenal quality, in that the content will mirror the quality (picking out instances of the quality in all epistemic possibilities), and in that across a wide range of nearby conceptually possible cases in which the underlying quality is varied while background properties are held constant, the content will covary to mirror the quality. Let us call this sort of phenomenal concept a *direct phenomenal concept*.

Not all experiences are accompanied by corresponding direct phenomenal concepts. Many of our experiences appear to pass without our forming any beliefs about them, and without the sort of concept formation that occurs in the Mary case. The clearest cases of direct phenomenal concepts arise when a subject attends to the quality of an experience, and forms a concept wholly based on the attention to the quality, "taking up" the quality into the concept. This sort of concept formation can occur with visual experiences, as in the Mary case, but it can equally occur with all sorts of other experiences: auditory and other perceptual experiences, bodily sensations, emotional experiences, and so on. In each case we can imagine the analog of Mary having such an experience for the first time, attending to it, and coming to have a concept of what it is like to have it. There is no reason to suppose that this sort of concept formation is restricted to entirely novel experiences. I can experience a particular shade of phenomenal redness for the hundredth time, attend to it, and form a concept of what it is like to have that experience, a concept whose content is based entirely on the character of the experience.

Direct phenomenal concepts can be deployed in a wide variety of beliefs, and other propositional attitudes. When Mary attends to her phenomenally red experience and forms her direct phenomenal concept *R*, she is thereby in a position to believe that tomatoes cause *R* experiences, to believe that others have *R* experiences, to believe that she previously had no *R* experiences, to desire more *R* experiences, and so on.

Perhaps the most crucial sort of deployment of a direct phenomenal concept occurs when a subject predicates the concept of the very experience responsible for constituting its content. Mary has a phenomenally red experience, attends to it and forms the direct phenomenal concept *R*, and forms the belief *This experience is R*, demonstrating the phenomenally red experience in question. We can call this special sort of belief a *direct phenomenal belief*.

We can also cast this idea within an experience-free ontology of qualities. In this framework, we can say that a direct phenomenal concept is formed by attending to a quality and taking up that quality into a concept whose content mirrors the quality, picking out instances of the quality in all epistemic possibilities. A direct phenomenal belief is formed when the referent of this direct phenomenal concept is identified with the referent of a corresponding demonstrative phenomenal concept, e.g. when Mary forms the belief that *This quality is R*. The general form of a direct phenomenal belief in this framework is $E=R$, where *E* is a demonstrative phenomenal concept and *R* is the corresponding direct phenomenal concept.

3.2 Some notes on direct phenomenal beliefs

(1) For a direct phenomenal belief, it is required that the demonstrative and direct concepts involved be appropriately "aligned". Say that Mary experiences phenomenal redness in both the left and right halves of her visual field, forms a direct phenomenal concept R based on her attention to the left half, forms a demonstrative concept of phenomenal redness based on her attention to the right half, and identifies the two by a belief of the form $E=R$. Then this is not a direct phenomenal belief, even though the same quality (phenomenal redness) is referred to on both sides, since the concepts are grounded in different instances of that quality. The belief has the right sort of content, but it does not have the right sort of constitution.

To characterize the required alignment more carefully we can note that all direct phenomenal concepts, like all demonstrative phenomenal concepts, are based in acts of attention to instances of phenomenal qualities. A direct phenomenal concept such as R does not characterize a quality *as* an object of attention, but it nevertheless requires attention to a quality for its formation. The same act of attention can also be used to form a demonstrative phenomenal concept E . A direct phenomenal belief (in the quality-based framework) will be a belief of the form $E=R$ where the demonstrative phenomenal concept E and the direct phenomenal concept R are *aligned*: that is, where they are based in the same act of attention.

One can simplify the language by regarding the act of attention as a demonstration. We can then say that both demonstrative and direct phenomenal concepts are based in demonstrations, and that a direct phenomenal belief is a belief of the form $E=R$ where the two concepts are based in the same demonstration.[*]

*[[Gertler (2001) has independently developed a related account of phenomenal introspection, according to which a phenomenal state is introspected when it is "embedded" in another state, and when the second state constitutes demonstrative attention to the relevant content by virtue of this embedding. On my account, things are the other way around: any "embedding" holds in virtue of demonstrative attention, rather than the reverse.]]

(2) As with all acts of demonstration and attention, phenomenal demonstration and attention involves a cognitive element. Reference to a phenomenal quality is determined in part by cognitive elements of a demonstration. These cognitive elements will also enter into determining the content of a corresponding direct phenomenal concept.

Consider two individuals with identical visual experiences. These individuals might engage in different acts of demonstration - e.g. one might demonstrate a red quality experienced in the right half of the visual field, and the other a green quality experienced in the left half of the visual field - and thus form distinct direct phenomenal concepts. Or they might attend to the same location in the visual field, but demonstrate distinct qualities associated with that location: e.g. one might demonstrate a highly specific shade of phenomenal redness, and the other a less specific shade, again resulting in distinct direct phenomenal concepts. These differences will be due to differences in the cognitive backgrounds of the

demonstrations in the two individuals. I will be neutral here about whether such cognitive differences are themselves constituted by underlying functioning, aspects of cognitive phenomenology, or both.

One can imagine varying the visual experiences and the cognitive background here independently. Varying visual experiences might yield a range of cases in which direct phenomenal concepts of phenomenal redness, greenness, and other hues are formed. Varying the cognitive background might yield a range of cases in which direct phenomenal concepts of different degrees of specificity (for example) are formed.

Along with this cognitive element comes the possibility of failed demonstration, if the cognitive element and the targeted experiential elements mismatch sufficiently. Take Nancy, who attends to a patch of phenomenal color, acting cognitively as if to demonstrate a highly specific phenomenal shade. Nancy has not attended sufficiently closely to notice that the patch has a nonuniform phenomenal color: let us say it is a veridical experience of a square colored with different shades of red on its left and right side.[*] In such a case, the demonstrative phenomenal concept will presumably refer to no quality at all: given its cognitive structure, it could refer only to a specific quality, but it would break symmetry for it to refer to either instantiated quality, and presumably uninstantiated qualities cannot be demonstrated.

*[[This sort of case was suggested to me by Delia Graff and Mark Johnston.]]

What of any associated direct phenomenal concept? It is not out of the question that the subject forms *some* substantive concept where a direct phenomenal concept would ordinarily be formed; perhaps a concept of an intermediate uninstantiated shade of phenomenal red, at least if the instantiated shades are not too different. Like a direct phenomenal concept, this concept will have a content that depends constitutively on associated qualities of experience (Inverted Nancy might form a concept of an intermediate phenomenal green), but it will not truly be a direct phenomenal concept, since its content will not directly mirror an underlying quality.

This possibility of cognitive mismatch affects the path from demonstration to a demonstrated phenomenal quality, but given that a phenomenal quality is truly demonstrated, it does not seem to affect the path from demonstrated phenomenal quality to a direct phenomenal concept. That is, as long as a phenomenal quality is demonstrated, and the cognitive act typical of forming a direct phenomenal concept based on such a demonstration is present, a direct phenomenal concept will be formed.

We might call a concept that shares the cognitive structure of a direct phenomenal concept a *quasi-direct* phenomenal concept; and we can call a belief with the same cognitive structure as a direct phenomenal belief a *quasi-direct* phenomenal belief. Like a direct phenomenal concept, a quasi-direct phenomenal concept arises from an act of (intended) demonstration, along with a characteristic sort of cognitive act. Unlike a direct phenomenal concept, a quasi-direct phenomenal concept is not required to have a content that is constituted by an underlying quality. Nancy's concept above is a quasi-direct phenomenal concept but not a direct phenomenal concept, for example.

We can call a quasi-direct phenomenal concept that is not a direct phenomenal concept a *pseudo-direct* phenomenal concept, and we can define a pseudo-direct phenomenal belief similarly. If the suggestion above is correct, then the only pseudo-direct phenomenal concepts are like Nancy's, in involving an unsuccessful demonstration. As long as a quasi-direct phenomenal concept is grounded in a successful demonstration, it will be a direct phenomenal concept. I will return to this claim later.

(3) All direct phenomenal concepts are pure phenomenal concepts, but not all pure phenomenal concepts are direct phenomenal concepts. To see this, note that Mary may well retain some knowledge of what it is like to see tomatoes even after she goes back into her black-and-white room, or while she shuts her eyes, or while she looks at green grass. She still has a concept of phenomenal redness that can be deployed in various beliefs, with the sort of epistemic relations to relational and demonstrative phenomenal concepts that is characteristic of pure phenomenal concepts. Inverted Mary (still Mary's physical twin) has a corresponding concept deployed in corresponding beliefs that *differ* in content from Mary's. And as before their corresponding beliefs *differ* in epistemic content, including and excluding different classes of epistemic possibilities. As before, Mary's concept is a concept of phenomenal redness *as* the quality it is, based on a lucid understanding of that quality, rather than on a mere relational or demonstrative identification. So as before, it is a pure phenomenal concept. But it is not a direct phenomenal concept, since there is no corresponding experience (or instantiated quality) that is being attended to or taken up into the concept. We can call this sort of concept a *standing* phenomenal concept, since it may persist in a way that direct phenomenal concepts do not.

There are some differences in character between direct and standing phenomenal concepts. Direct phenomenal concepts may be very fine-grained, picking out a very specific phenomenal quality (a highly specific shade of phenomenal redness, for example). Standing phenomenal concepts are usually more coarse-grained, picking out less specific qualities. One can note this phenomenologically from the difficulty of "holding" in mind specific qualities as opposed to coarser categories when relevant visual experiences are not present; and this is also brought out by empirical results showing the difficulty of reidentifying specific qualities over time.[*] It usually seems possible for a direct phenomenal concept to yield a corresponding standing phenomenal concept as a "successor" concept once the experience in question disappears, at the cost of some degree of coarse-graining.

*[[See Raffman 1995 for a discussion of these results in an argument for an antirepresentationalist "presentational" analysis of phenomenal concepts that is very much compatible with the analysis here.]]

As with direct phenomenal concepts, the content of standing phenomenal concepts does not conceptually supervene on the physical (witness Mary and Inverted Mary, back in their rooms). A question arises as to what determines their content. I will not try to analyze that matter here, but I think it is plausible that their content is determined by some combination of (i) nonsensory phenomenal states of a cognitive sort, which bear a relevant relation to the original phenomenal quality in question - e.g. a faint Humean phenomenal "idea" that is relevantly related to the original "impression"; (ii) dispositions to have such states; and (iii) dispositions to recognize instances of the phenomenal quality in question. It is not implausible that Mary and Inverted Mary (back in their rooms) still differ in some or all of these respects, and that these respects are constitutively responsible for the difference in the content of their concepts.

One might be tempted to use the existence of standing phenomenal concepts to argue against the earlier analysis of direct phenomenal concepts (that is, of concepts akin to those Mary acquires on first experiencing phenomenal redness) as constituted by the quality of the relevant instantiated experience. Why not assimilate them to standing phenomenal concepts instead, giving a unified account of the two? In response, note first that it remains difficult to conceive of the content of direct phenomenal concepts varying independently of the phenomenal quality in question, whereas it does not seem so difficult to conceive of the content of standing phenomenal concepts varying independently. And second, note that the difference in specificity between direct and standing phenomenal concepts gives some reason to believe that they are constituted in different ways.

The lifetime of a direct phenomenal concept is limited to the lifetime of the experience (or the instantiated quality) that constitutes it. (In some cases a specific phenomenal concept might persist for a few moments due to the persistence of a vivid iconic memory, but even this will soon disappear.) Some might worry that this lack of persistence suggests that it is not a concept at all, since concepthood requires persistence. This seems misguided, however: it is surely possible for a concept to be formed moments before a subject dies. The concepts in question are still predicable of any number of entities, during their limited lifetimes, and these predications can be true or false. (E.g., Mary may falsely believe that her sister is currently experiencing R.) This sort of predicability, with assessability for truth or falsehood, seems sufficient for concepthood; at least it is sufficient for the uses of concepthood that will be required here.

(4) As with pure phenomenal concepts generally, we do not have public language expressions that distinctively express the content of direct phenomenal concepts. Public reference to phenomenal qualities is always fixed relationally, it seems: by virtue of a relation to certain external stimuli, or certain sorts of behavior, or certain demonstrations. (As Ryle said, there are no "neat" sensation words.) Of course Mary can express a pure phenomenal concept by introducing her own term, such as 'R', or by using an old term, such as 'red', with this stipulated meaning. But this use will not be public, at least in the limited sense that there is no method by which we can ensure that other members of the community will use the term with the same epistemic content. One can at best ensure that they pick out the same quality by picking it out under a different epistemic content (e.g. as the quality Mary is having at a certain time), or by referring through semantic deference (as the quality that Mary picks out with 'R'.) In this sense it seems that any resulting language will be "private": it can be used with full competence by just one subject, and others can use it only deferentially.

(An exception may arguably be made for terms expressing *structural* pure phenomenal concepts - e.g. phenomenal similarity and difference and perhaps phenomenal spatial relations - which arguably do not rely on relational reference-fixing.)

Of course the view I have set out here is just the sort of view that Wittgenstein directed his "private language" argument against. The nature of the private language argument is contested, so in response I can say only that I have seen no reconstruction of it that provides a strong case against the view I have laid out. Some versions of the argument seem to fall prey to the mistake just outlined, that of requiring a

strong sort of "repeatability" for concept possession (and an exceptionally strong sort at that, requiring the recognizability of correct repeated application). A certain sort of repeatability is required for concept possession, but it is merely the "hypothetical repeatability" involved in *present* predicability of the concept to actual and hypothetical cases, with associated truth-conditions. Another reconstruction of the argument, Kripke's, provides no distinctive traction against my analysis of direct phenomenal concepts: any force that it has applies to concepts quite generally.

(One might even argue that Kripke's argument provides *less* traction in the case of direct phenomenal concepts, as this is precisely a case in which we can see how a determinate application-condition can be constituted by an underlying phenomenal quality. Kripke's remarks (pp. 41-51) about associated phenomenal qualities - e.g. a certain sort of "headache" - being irrelevant to the content of concepts such as addition apply much less strongly where *phenomenal* concepts are concerned. Of course there is more to say here, but in any case it is a curiosity of Kripke's reconstruction of the argument that it applies least obviously to the phenomena at which Wittgenstein's argument is often taken to be aimed.)

4 The Epistemology of Phenomenal Belief

4.1 Incorrigibility

A traditional thesis in the epistemology of mind is that first-person beliefs about phenomenal states are *incorrigible*, or *infallible* (I use these terms equivalently), in that they cannot be false. In recent years such a thesis has been widely rejected. This rejection stems from both general philosophical reasoning (e.g., the suggestion that if beliefs and experiences are distinct existences, there can be no necessary connection between them) and from apparent counterexamples (e.g. a case where someone, expecting to be burnt, momentarily misclassified a cold sensation as hot). In this light, it is interesting to note that the framework outlined in this paper so far supports an incorrigibility thesis, albeit a very limited one.

Incorrigibility Thesis: A direct phenomenal belief cannot be false.

The truth of this thesis is an immediate consequence of the definition of direct phenomenal belief. A direct phenomenal concept by its nature picks out instances of an underlying demonstrated phenomenal quality, and a direct phenomenal belief identifies the referent of that concept with the very demonstrated quality (or predicates the concept of the very experience that instantiated the quality), so its truth is guaranteed.

If we combine this thesis (which is more or less true by definition) with the substantive thesis that there are direct phenomenal beliefs (which is argued earlier in this paper), then we have a substantive incorrigibility thesis, one that applies to a significant range of actual beliefs.[*]

*[[Pollock (1986, pp. 32-33) entertains a version of this sort of view as a way of supporting incorrigibility, discussing a "Containment Thesis" according to which experiences are constituents of beliefs about experiences. He rejects the view on the grounds that (1) it does not support incorrigibility of negative beliefs about experiences (e.g. the belief that one is not

having a given experience), which he holds to be required for incorrigibility in general, and that (2) that having an experience does not suffice to have the relevant belief, so having the belief also requires thinking about the experience, which renders the incorrigibility thesis trivial. I discuss both of these points below.]]

The thesis nevertheless has a number of significant limitations. The first is that most phenomenal beliefs are not direct phenomenal beliefs, so most phenomenal beliefs are still corrigible. The most common sort of phenomenal belief arguably involves the application of a *pre-existing* phenomenal concept (either a relational phenomenal concept or a standing pure phenomenal concept) to a new situation, as with the beliefs typically expressed by claims such as 'I am having a red experience' or 'I am in pain'. These are not direct phenomenal beliefs, and are almost certainly corrigible.

There are also cases in which a direct phenomenal concept is applied to a quality (or an experience) other than the one that constituted it, as when one forms a direct phenomenal concept *R* based on a quality instantiated in the left half of one's visual field, and applies it to a quality instantiated in the right half. These are also not direct phenomenal beliefs, and are again almost certainly corrigible.

(The second sort of case brings out a further limitation in the incorrigibility thesis: it does not yield incorrigibility in virtue of content. If the left and right qualities in the case above are in fact the same, then the resulting nondirect phenomenal belief will arguably have the same content as the corresponding direct phenomenal belief, but the incorrigibility thesis will not apply to it. The domain of the incorrigibility thesis is constrained not just by content, but by underlying constitution.)

It is plausible that all the standard counterexamples to incorrigibility theses fall into classes like these, particularly the first. All the standard counterexamples appear to involve the application of pre-existing phenomenal concepts (*pain*, *hot*, *red experience*). So none of the standard counterexamples apply to the incorrigibility thesis articulated here.

There is a natural temptation to find further counterexamples to the incorrigibility thesis. For example, one might consider a case in which a subject's experience changes very rapidly, and argue that the corresponding direct phenomenal concept must lag behind. In response to these attempted counterexamples, the most obvious reply is that these cannot truly be counterexamples, since the truth of the incorrigibility thesis is guaranteed by the definition of direct phenomenal belief. If the cases work as described, they do not involve direct phenomenal beliefs: they either involve a concept that is not a direct phenomenal concept, or they involve a direct phenomenal concept predicated of a quality other than the one that constitutes it. At best, they involve what I earlier called pseudo-direct phenomenal beliefs: beliefs that share the cognitive structure of a direct phenomenal beliefs (and thus are quasi-direct phenomenal beliefs) but that are not direct phenomenal beliefs.

One need not let matters rest there, however. I think that these counterexamples can usually be analyzed away on their own terms, so that the purported pseudo-direct phenomenal beliefs in question can be seen as direct phenomenal beliefs, and as correct. In the case of a rapidly changing experience, one can plausibly hold that the content of a direct phenomenal concept covaries immediately with the underlying quality, so that there is no moment at which the belief is false. This is just what we would expect, given

the constitutive relation suggested earlier. We might picture this schematically by suggesting that the basis for a direct phenomenal concept contains within it a "slot" for an instantiated quality, such that the quality that fills the slot constitutes the content. In a case where experience changes rapidly, the filler of the slot changes rapidly, and so does the content.

Something similar goes for many other examples involving quasi-direct phenomenal beliefs. Take a case where a subject attends to two different visual qualities (demonstrating them as $E1$ and $E2$), and mistakenly accepts $E1=E2$. In this case, someone might suggest that if the subject forms specific quasi-direct phenomenal concepts $R1$ and $R2$ based on the two acts of attention, these must have the same content, leading to false quasi-direct phenomenal beliefs (and thus to pseudo-direct phenomenal beliefs). But on my account, this case is better classified as one in which $R1$ and $R2$ are direct phenomenal concepts with different contents, yielding two correct direct phenomenal beliefs $E1=R1$ and $E2=R2$. The false beliefs here are of the form $E1=R2$, $E2=R1$, and $R1=R2$. The last of these illustrates the important point that identities involving two direct phenomenal concepts, like identities involving two pure phenomenal concepts more generally, are not incorrigible.

Other cases of misclassification can be treated similarly. In the case in which a subject expecting to be burnt misclassifies a cold sensation as hot, someone might suggest that any quasi-direct phenomenal concept will be a concept of phenomenal hotness, not coldness. But one can plausibly hold that if a quasi-direct phenomenal concept is formed, it will be a concept of phenomenal coldness and will yield a correct direct phenomenal beliefs. The subject's mistake involves misclassifying the experience under standing phenomenal concepts, and perhaps a mistaken identity involving a direct and a standing phenomenal concept.

It is arguable that most cases involving quasi-direct phenomenal beliefs can be treated this way. The only clear exceptions are cases such as Nancy's, in which no phenomenal quality is demonstrated and so no substantive direct phenomenal concept is formed. It remains plausible that as long as a quality is demonstrated, the cognitive act in question will yield a direct phenomenal concept with the right content, and a true direct phenomenal belief. If that is correct, one can then accept a broader incorrigibility thesis applying to any quasi-direct phenomenal belief that is based in a successful demonstration of a phenomenal quality. I will not try to establish this thesis conclusively, since I will not need it, and since the incorrigibility thesis for direct phenomenal beliefs is unthreatened either way. But it is interesting to see that it can be defended.

One might suggest that the incorrigibility thesis articulated here (in either the narrower or the broader version) captures the *plausible core* of traditional incorrigibility theses. A number of philosophers have had the sense that there is something correct about the incorrigibility theses, which is not touched by the counterexamples. This is reflected, for example, in Chisholm's distinction between "comparative" and "noncomparative" uses of "appears" talk, where only the noncomparative uses are held to be incorrigible. I think that this is not quite the right distinction: even noncomparative uses can be corrigible, when they correspond to uses of pure phenomenal concepts outside direct phenomenal beliefs. But perhaps a thesis restricted to direct phenomenal beliefs might play this role.

Certainly the analysis of direct phenomenal beliefs shows why the most common general philosophical argument against incorrigibility does not apply across the board. In the case of direct phenomenal beliefs, beliefs and experiences are *not* entirely distinct existences. It is precisely because of the constitutive connection between experiential quality and belief that the two can be necessarily connected.

Another limitation: Sometimes incorrigibility theses are articulated in a "reverse" or bidirectional form, holding that all phenomenal states are incorrigibly known, or at least incorrigibly knowable. Such a thesis is not supported by the current discussion. Most phenomenal states are not attended to, are not taken up into direct phenomenal concepts, and so are not the subjects of direct phenomenal beliefs. And for all I have said, it may be that some phenomenal states, such as fleeting or background phenomenal states, *cannot* be taken up into a direct phenomenal concept, perhaps because they cannot be subject to the right sort of attention. If so, they are not even incorrigibly knowable, let alone incorrigibly known.

Incorrigibility theses are also sometimes articulated in a "negative" form, requiring that a subject cannot be mistaken in their belief that they are *not* having a given sort of experience. No direct phenomenal belief is a negative phenomenal belief, so the current framework does not support this thesis, and I think the thesis is false in general.

A final limitation: Although direct phenomenal beliefs are incorrigible, subjects are not incorrigible about whether they are having a direct phenomenal belief. For example, if I am not thinking clearly, I might misclassify a belief involving a standing phenomenal concept as a direct phenomenal belief. And in the Nancy case above, if Nancy is philosophically sophisticated she might well think that she is having a direct phenomenal belief, although she is not.

One could argue that this lack of higher-order incorrigibility prevents the first-order incorrigibility thesis from doing significant epistemological work. The matter is delicate: higher-order incorrigibility is probably too strong a requirement for an epistemologically useful incorrigibility thesis. But on the other side, *some* sort of further condition is required for a useful thesis. For example, any member of the class of true mathematical beliefs is incorrigible (since it is necessarily true), but this is of little epistemic use to a subject who cannot antecedently distinguish true and false mathematical beliefs. A natural suggestion is that some sort of higher-order accessibility is required.

Intermediate accessibility requirements might include something like the following. For the incorrigibility of a direct phenomenal belief to be epistemologically significant, a subject must know that it is a direct phenomenal belief, or at least be justified in so believing; or a subject must be capable of so knowing on reflection; or direct phenomenal beliefs must be cognitively or phenomenologically distinctive as a class relative to nondirect phenomenal beliefs.

I am sympathetic with the sufficiency of a thesis like the last, if properly articulated. Its truth seems to turn on questions about quasi-direct and pseudo-direct phenomenal beliefs. If there are many pseudo-direct phenomenal beliefs, and if there is nothing cognitively or phenomenologically distinctive about direct phenomenal beliefs by comparison, then direct phenomenal beliefs will simply be distinguished as

quasi-direct phenomenal beliefs with the right sort of content, and the incorrigibility claim will be relatively trivial. On the other hand, if pseudo-direct phenomenal beliefs are rare, or if direct phenomenal beliefs are a cognitively or phenomenologically distinctive subclass, then it is more likely that incorrigibility will be nontrivial and carry epistemological significance.

If pseudo-direct phenomenal beliefs are restricted to cases in which no phenomenal quality is demonstrated, like the case of Nancy (as I have suggested), then the incorrigibility thesis will hold of a class of beliefs that can be distinctively and independently characterized in cognitive and phenomenological terms: the class of quasi-direct phenomenal beliefs which are based in a successful demonstration. This would render the incorrigibility claim entirely nontrivial, and it would make it more likely that it could do epistemological work. But I will not try to settle this matter decisively here, and I will not put the incorrigibility thesis to any epistemological work in what follows.

It might be thought that the incorrigibility thesis suffers from another problem: that direct phenomenal beliefs are incorrigible because they are *trivial*. After all, beliefs such as *I am here* or *This is this* are (close to) incorrigible, but only because they are (almost) trivial. (The qualifications are present because of the the arguable nontriviality of my existence and spatial locatedness in one case, and because of the possibility of reference failure for the demonstrative in the other.)

The analogy fails, however. The trivial beliefs in question are (almost) cognitively insignificant: they are (almost) a priori, containing (almost) no cognitively significant knowledge about the world. This is reflected in the fact that they hardly constrain the class of a priori epistemic possibilities: they are true of (almost) all such possibilities, considered as hypotheses about the actual world. (Two-dimensionally: these beliefs have a epistemic intension that is (almost) conceptually necessary.) A direct phenomenal belief, by contrast, is cognitively significant: it heavily constrains the class of a priori epistemic possibilities, and is false in most of them (considered as actual). For example, Mary's direct phenomenal belief, on leaving her room, is false of all worlds (considered as actual) in which the subject is not experiencing phenomenal redness. (Two-dimensionally: The epistemic intension of a direct phenomenal belief is conceptually contingent.) So direct phenomenal beliefs, unlike the beliefs above, are entirely nontrivial.

So: the incorrigibility thesis articulated here has a number of limitations, but it nevertheless applies to a significant class of nontrivial phenomenal beliefs.

4.2 Acquaintance and justification

At this point it is natural to ask: if we can form this special class of incorrigible, distinctively constituted beliefs where phenomenal states and properties are concerned, why cannot we do so where other states and properties are concerned? Why cannot we form direct height concepts, for example, whose epistemic content is directly constituted by our height properties, and which can be deployed in incorrigible direct height beliefs? Or similarly for direct chemical beliefs, direct age beliefs, direct color beliefs, and so on?

At one level, the answer is that we simply cannot. If one tries to form a direct height concept - one whose content depends constitutively on an instantiated height - the best one can do is form a relational height concept (*my height, the height of my house*) or a demonstrative height concept. But these are not pure height concepts at all. They are analogous only to *red_C* or *E*, in that their subjunctive content may depend on the property in question but their epistemic content does not.

It is arguable whether pure height concepts exist at all: that is, whether there is any concept whose epistemic content picks out a certain height (say, two meters) in any epistemic possibility. But even if there are pure height concepts, they are not direct height concepts. Perhaps one can independently form a pure height concept of a given height (two meters), which might coincide with an instantiated height, but it will not depend constitutively on an instantiated height. The best one can do is attend to an object, have an experience or judgment concerning its height, and use this experience or judgment as the epistemic content of a "pure" height concept. But here the instantiated height property is not constitutively relevant to the concept's content, but only causally relevant: it is the height experience or judgment that is constitutively relevant, and the experience or judgment is only causally dependent on the height. In no case does the epistemic content of a height concept depend constitutively on a demonstrated height property, or on any instantiated height property at all.

Proponents of certain direct realist views may hold that it is possible to form a direct concept of a height property (or other perceivable external properties), by demonstrating it and taking it up into a concept in a manner analogous to the manner suggested for phenomenal properties. I think that this is implausible. In a case where an object is two meters tall but appears to be one meter tall, any "pure" height concept formed as a result will be a concept of one meter, not of two meters. There may be a demonstrative concept of two meters, but that is not enough. More generally, considering a range of cases in which height and experience are varied independently, we can see that any contribution of the height to a pure concept is "screened off" by the contribution of the experience. This suggests that if anything is playing a constitutive role in the concept's content, it is the experience and not the external property.[*]

*[[There may be further moves available to the direct realist. For example, a direct realist might hold that the constitutive role of external properties is restricted to cases of veridical perception, and that nonveridical perception must be treated differently. I think that this sort of restriction threatens to trivialize the constitution thesis, as any causal connection might be seen as a "constitutive" connection by a relevantly similar restriction. (If A causes B which necessitates C, then A is contingently connected to C; but if we restrict attention to cases where A causes B, then A necessitates C relative to this restriction.) And the case remains formally disanalogous to the case of direct phenomenal concepts, in which there is no factor distinct from the quality that even looks like it screens off the contribution of the quality to the concept. But there is undoubtedly more to say here. In what follows I will assume that the direct realist view is incorrect, but direct realists are free to hold that what I say about phenomenal properties applies equally to the relevant external properties.]]

The same goes for chemical concepts, age concepts, and external color concepts. Although we can form many such concepts, in no case is it possible to form a direct concept: that is, a concept whose epistemic content depends constitutively on a demonstrated property. It seems that only phenomenal properties can support direct concepts.

This conclusion is apparently revealed by an examination of cases; but it would be preferable not to leave

it as a brute conclusion. In particular, it is natural to suggest that the conclusion holds because we bear a special relation to the phenomenal properties instantiated in our experience: a relation that we do not bear to the other instantiated properties in question, and a relation that is required in order to form a direct concept of a property in the manner described. This relation would seem to be a peculiarly intimate one, made possible by the fact that experiences lie at the heart of the mind rather than standing at a distance from it; and it seems to be a relation that carries the potential for conceptual and epistemic consequences. We might call this relation *acquaintance*.

As things stand, acquaintance has been characterized only as that relation between subjects and properties that makes possible the formation of direct phenomenal concepts; so it is not yet doing much explanatory work. But having inferred the relation of acquaintance, we can put it to work. As characterized, acquaintance is a relation that makes possible the formation of pure phenomenal concepts, and we have seen that pure phenomenal concepts embody a certain sort of lucid understanding of phenomenal properties. So acquaintance is a relation that makes this sort of lucid understanding possible. As such, it is natural to suppose that the relation can also do work in the epistemic domain. If so, the result will be an attractive picture in which the distinctive conceptual character and the distinctive epistemic character of the phenomenal domain have a common source.

It is independently plausible to hold that phenomenal properties and beliefs have a distinctive epistemic character. Many have held that phenomenal properties can (at least sometimes) be known with a distinctive sort of justification, or even with certainty; and many have held that phenomenal beliefs have a special epistemic status. Even those who explicitly deny this will often tacitly concede that there is at least a *prima facie* case for this status: for example, it is striking that those who construct skeptical scenarios almost always ensure that that phenomenal properties are preserved. So it is arguable that simply having a phenomenal property provides the potential for a strong sort of phenomenal knowledge. Something similar is suggested by the Mary case: Mary's experience of the phenomenal property R allows her to have not just a distinctive phenomenal belief, but also distinctive phenomenal knowledge. Some element of this distinctive epistemic character can be captured in the present framework.

One natural suggestion is the following: direct phenomenal beliefs are always justified. Certainly Mary's belief on leaving her room seems to be justified, and most other examples seem to fit this thesis. This thesis has to be modified slightly. There are presumably subjects who are so irrational or confused that none of their beliefs qualify as justified, so that their direct phenomenal beliefs are not justified either. And perhaps there could be subjects who are so confused about phenomenology that they accept not just direct phenomenal beliefs but their negations, casting doubt on whether either belief is truly justified. To meet this sort of case, we might adjust the thesis to say that all direct phenomenal beliefs have some *prima facie* justification, where *prima facie* justification is an element of justification that can sometimes be overridden by other elements, rendering a belief below the threshold for "justification" simpliciter. Something similar presumably applies to other features of a belief that might seem to confer justification, such as being inferred from justified beliefs by a justified rule of inference.

Assuming that something like this is right: it is nevertheless one thing to make the case that direct phenomenal beliefs are (*prima facie*) justified, and another to give an account of what this justification

consists in. It may be tempting to appeal to incorrigibility; but incorrigibility alone does not entail justification (as the mathematical case shows), and while certain higher-order accessibility theses might close the gap, it is not obvious that they are satisfied for direct phenomenal beliefs.

A better idea is to appeal to the acquaintance relation, thus unifying the distinctive conceptual and epistemic character of phenomenal beliefs. In particular, one might assert the following:

Justification Thesis: When a subject forms a direct phenomenal belief based on a phenomenal quality, then that belief is *prima facie* justified by virtue of the subject's acquaintance with that quality.

Certainly many philosophers, including especially sense-data theorists and more recent foundationalists, have appealed to a relation of acquaintance (or "direct awareness") in supporting the special epistemic status of phenomenal beliefs. The current account offers a more constrained version of such a thesis, suggesting that it holds for a special class of phenomenal beliefs (on which the epistemic content of a predicated concept is required to mirror and be constituted by the acquainted quality, to which it is applied), and on the basis of a relation whose existence we have made an independent case for.

Some philosophers (e.g. Russell 1910; Fumerton 1995) have held that we are "acquainted with acquaintance", and have made the case of its existence that way. I think there is something to the idea that our special epistemic relation to experience is revealed in our experience, but I note that the proponent of acquaintance is not forced to rely on such a thesis. It is equally possible to regard acquaintance as a theoretical notion, inferred to give a unified account of the distinctive conceptual and epistemic character that we have reason to believe is present in the phenomenal domain.

Acquaintance can be regarded as a basic sort of epistemic relation between a subject and a property. Most fundamentally, it might be seen as a relation between a subject and an *instance* of a property: I am most directly acquainted with *this instance* of phenomenal greenness. This acquaintance with an instance can then be seen to confer a derivative relation to the property itself. Or in the experience-based framework, one might regard acquaintance as most fundamentally a relation between a subject and an experience, which confers a derivative relation between the subject and the phenomenal properties of the experience. But I will usually abstract away from these fine details. What is central will be the shared feature that whenever a subject has a phenomenal property, the subject is acquainted with that phenomenal property.

Even if acquaintance is a theoretical notion, it clearly gains some pretheoretical support from the intuitive view that beliefs can be epistemically grounded in experiences, where experiences are not themselves beliefs but nevertheless have an epistemic status that can help justify a belief. One might view acquaintance as capturing that epistemic status.

In certain respects (though not in all respects), the justification of a direct phenomenal belief by an experience can be seen as analogous to the justification of an inferred belief by another belief. For an inferred belief to be *prima facie* justified, there are three central requirements: one concerning the content of the belief in relation to the justifying state, one concerning the natural connection between the belief

and the justifying state, and one concerning the epistemic status of the justifying state. First, the epistemic content of the belief must be appropriately related to that of the belief that it is inferred from. Second, the belief must be appropriately be caused by the justifying belief. Third, the justifying belief must itself be justified.

In the *prima facie* justification of a direct phenomenal belief by an experience, there are three factors of the same sort. First, content: the epistemic content of the direct phenomenal belief must mirror the quality of the experience. Second, a natural connection: the phenomenal belief must be appropriately constituted by the experience. And third, epistemic status: the subject must be acquainted with the justifying quality. The details of the requirements are different, as befits the difference between belief and experience, but the basic pattern is very similar.

It is plausible that a subject can have phenomenal properties without having corresponding concepts, or corresponding beliefs, or corresponding justification.[*] If so, the same goes for acquaintance. Acquaintance is not itself a conceptual relation: rather, it makes certain sorts of concepts possible. And it is not itself a justificatory relation: rather, it makes certain sorts of justification possible. Phenomenal concepts and phenomenal knowledge require not just acquaintance, but acquaintance in the right cognitive background: a cognitive background that minimally involves a certain sort of attention to the phenomenal quality in question, a cognitive act of concept formation, the absence of certain sorts of confusion and other undermining factors (for full justification), and so on. But it is acquaintance with the quality or the experience itself that does the crucial justifying work.

*[[Nothing I have said so far requires that experiences can exist without concepts; at most, it requires that experiences can exist without phenomenal concepts. So what I have said may be compatible with views on which experiences depend on other concepts in turn. Still, I think it is independently plausible that experiences do not require concepts for their existence, and I will occasionally assume this in what follows. This is not to deny that *some* experiences depend on concepts, and it is also not to deny that experiences have representational content.

My own view is that at least for perceptual experiences (and perhaps for all experiences), experiences have representational content by virtue of their phenomenology, where this content is sometimes conceptual and sometimes nonconceptual. This yields an interesting possibility (developed in forthcoming work): the constitutive relation between phenomenal states and phenomenal concepts might be extended to yield a similar constitutive relation between perceptual phenomenal states and a special class of perceptual concepts, by virtue of the phenomenal states' representational content. Such an account might yield some insight into the content and epistemology of perceptual belief.]]

Some philosophers hold that only a belief can justify another belief. It is unclear why this view should be accepted. The view has no pretheoretical support: pretheoretically, it is extremely plausible that experiences (e.g. a certain experience of phenomenal greenness) play a role in justifying beliefs (e.g. my belief that there is something green in front of me, or my belief that I am having a certain sort of experience), even though experiences are not themselves beliefs. And the view has no obvious theoretical support. Perhaps the central motivation for the view comes from the idea that inference is the only sort of justification that we understand and have a theoretical model for, and that we have no model for any other sort of justification. But this is obviously not a strong reason, and the account I have just sketched suggests a theoretical model of how experiences can justify beliefs that fits well with our pretheoretical

intuitions. So it seems that the cases of justification of beliefs by other beliefs and by experiences are on a par here.

Another motivation for the view comes from the thesis that for a state to justify another state, it must itself be justified (along with the claim that only beliefs can be justified. But again, it is unclear why this thesis should be accepted. Again, it is pretheoretically reasonable to accept that beliefs are justified by experiences, and that experiences are not themselves the sort of states that can be justified or unjustified. And there is no obvious theoretical reason to accept the thesis. It may be that for a state to justify, it must have *some* sort of epistemic status, but there is no clear reason why the status of acquaintance should be insufficient.

(BonJour (1978) suggests that the denial that justifying states must be justified is an ad hoc maneuver aimed at stopping the regress argument against foundationalism. But considerations about foundationalism and about regress arguments have played no role in my claims: the claims are independently supported by observations about the epistemic and conceptual relations between belief and experience. BonJour also claims that a justifying state must involve assertive content; but again, there is no clear pretheoretical or theoretical reason to accept this. Pretheoretically: experiences can justify beliefs without obviously involving assertive content. Theoretically: acquaintance with a property makes the property available to a subject in a manner that makes concepts and assertions involving the property possible, and that enables these assertions to be justified. There is no reason why this requires acquaintance to itself involve an assertion.)

A number of epistemological issues remain. One concerns the strength of the justification of phenomenal beliefs. It is often held that phenomenal beliefs are (or can be) *certain*, for example. Can the present framework deliver this? It can certainly deliver incorrigibility, but certainty requires something different. I think that the relevant sense of certainty involves something like *knowledge beyond skepticism*: intuitively, knowledge such that one's epistemic situation enables one to rule out all skeptical counterpossibilities. There is an intuition that phenomenal belief at least sometimes involves this sort of knowledge beyond skepticism, as the standard construction of skeptical scenarios suggests.

This epistemic status might be captured by a claim to the effect that acquaintance with a property enables one to eliminate all (a priori) epistemic possibilities in which the property is absent.[*] If so, then in the right cognitive background (with sufficient attention, concept formation, lack of confusion, and so on), the justification of a direct phenomenal belief *P* by acquaintance with a property will sometimes enable a subject not just to know that *P* by the usual standards of knowledge, but to eliminate all skeptical counterpossibilities in which *P* is false. This matter requires further exploration, but one can see at least the beginnings of a reasonable picture.[*]

*[[I argued in *The Conscious Mind* that something like acquaintance is required to secure certainty, and that a mere causal connection or reliable connection cannot do the job. If the justification of a belief is based solely on a reliable or causal connection, the subject will not be in a position to rule out skeptical scenarios in which the connection is absent and the belief is false, so the belief will not be certain. In response, a number of philosophers, including Bayne (2001), have argued that acquaintance accounts can be criticized in a similar way. Bayne notes that acquaintance alone is compatible with the

absence of certainty (e.g. in conditions of inattention), so certainty requires background factors in addition to acquaintance; but we cannot be certain that these factors obtain, so we cannot rule out skeptical scenarios in which they fail to obtain, so a phenomenal belief cannot be certain.

This argument stems from a natural misreading of my argument against reliabilist accounts. The argument is not: certainty requires certainty about the factors that enable certainty, and a reliabilist account cannot deliver this sort of certainty. That argument would require a strong version of a CJ thesis, that certain justification requires certainty about the basis of certain justification (analogous to the KJ thesis that justification requires knowledge of justification). I think such a thesis should clearly be rejected. The argument is rather: certainty about *P* requires (first-order) "knowledge beyond skepticism", or an epistemic state that enables a subject to rule out all skeptical scenarios in which *P* is false. Reliabilism by its nature cannot do this: there will always be skeptical scenarios in which the reliable connection fails and in which *P* is false.

Bayne's argument against acquaintance gives an analog of the invalid CJ argument. At most this establishes that we cannot rule out scenarios in which the belief is uncertain. Even this is unclear, as it is not obvious that certainty about certainty requires certainty about the factors enabling certainty. But even if this point is granted, the existence of skeptical scenarios in which the belief is uncertain does not entail the existence of skeptical scenarios in which *P* is false. Acquaintance yields certainty about experiences, not about beliefs: it enables one to directly rule out skeptical scenarios in which *P* is false, whether or not it enables one to rule out skeptical scenarios in which a belief is uncertain. In cases of justification by a reliable connection, there are separate reasons to hold that skeptical scenarios in which *P* is false cannot be ruled out, but in the case of acquaintance, these reasons do not apply. (Note: the published version of Bayne's article takes these points into account and offers some further considerations.)]]

A second further issue: Can the justification thesis be extended to all pure phenomenal concepts, including standing phenomenal concepts? There is some intuitive appeal in the idea that application of a standing phenomenal concept to an instantiated quality may also carry some justification by virtue of acquaintance with the quality (perhaps under the restriction that the content of the standing concept match the quality, and that there be an appropriate natural connection between the quality and the belief). If this belief were justified directly by acquaintance, however, we would need an account of justification by acquaintance that does not give a central role to constitution. Such an account is not out of the question, but it is worth noting that justification for beliefs involving standing phenomenal concepts can also be secured indirectly.

Indirect justification for such beliefs can be secured by virtue of the plausible claim that any belief of the form *S=R* is (prima facie) justified, where *S* and *R* are standing and direct phenomenal concepts with the same epistemic content. This is an instance of the more general claim that any belief of the form *A=B* is justified when *A* and *B* have the same epistemic content. (This thesis may need some restriction to handle cases of deep hyperintensionality, but it is plausibly applicable in this case.) Such beliefs are plausibly justified a priori: experience may enter into a grasp of the concepts involved in such a belief, but it does not enter into the belief's justification. If so, then beliefs involving standing phenomenal concepts will inherit justification by a priori inference from direct phenomenal beliefs, which will be justified in virtue of the Justification Thesis.

Finally, a note on ontology: Talk of acquaintance often brings sense-data theories to mind, so it may be worth noting that a commitment to phenomenal realism and to acquaintance does not entail a commitment to sense-data. First, the picture is entirely compatible with an "adverbial" subject-property

model, and with other quality-based ontologies on which there are phenomenal properties but not phenomenal individuals. Second, even if one accepts the existence of phenomenal individuals such as experiences, one might well reject a sense-data model of perception, on which one perceives the world by perceiving these entities.

It is also worth noting that one need not regard the acquaintance relation that a subject bears to a phenomenal property as something ontologically over and above the subject's instantiation of the property, requiring a subject-relation-quality ontology at the fundamental level. It is arguable that it is a conceptual truth that to have a phenomenal quality is to be acquainted with it (at least insofar as we have a concept of acquaintance that is not wholly theoretical). Certainly it is hard to conceive of a scenario in which a phenomenal quality is instantiated but no-one is acquainted with it. If so, then the picture I have sketched is combined with a simple subject-quality ontology, combined with this conceptual truth. The ontological ground of all this might lie in the nature of phenomenal qualities, rather than in some ontologically further relation.

4.3 Epistemological problems for phenomenal realism

Phenomenal realism, especially property dualism, is often thought to face epistemological problems. In particular, it is sometimes held that these views make it hard to see how phenomenal beliefs can be justified or can qualify as knowledge, since the views entail that phenomenal beliefs do not stand in the right sort of relationship to experiences. If what I have said so far is right, this cannot be correct. But it is worth looking at the arguments more closely.[*]

*[[I discussed these arguments at length in Chapter 5 of *The Conscious Mind*, on "The paradox of phenomenal judgment". I now think that discussion is at best suboptimal. The final section of the chapter (pp. 203-209) put forward a preliminary and sketchy version of the view of phenomenal concepts I have discussed in this paper, but I did not give it a central epistemological role (except in a tentative suggestion on pp. 207-8). I now think that this view of phenomenal concepts is central to the epistemology. So the discussion in this paper can be viewed in part as a replacement for that chapter.]]

The most influential arguments of this sort have been put forward by Sydney Shoemaker (1975). Shoemaker's arguments are intended as an argument against a view that admits the conceptual possibility of "absent qualia": a experience-free functional duplicate of an experiencing being. The view under attack is slightly stronger than phenomenal realism (a phenomenal realist could admit inverted qualia without absent qualia), is slightly weaker than a view on which zombies (experience-free physical duplicates) are conceptually possible, and is weaker than property dualism. But for the purposes of the argument, it will not hurt to assume a property dualist version of the view on which zombies are metaphysically possible. This has the effect of making Shoemaker's arguments harder to answer, not easier. The answers can easily be adapted to weaker versions of phenomenal realism.

The starkest version of Shoemaker's epistemological argument runs as follows:

- (1) If phenomenal realism is true, experiences are causally irrelevant to phenomenal

beliefs;

(2) If experiences are causally irrelevant to phenomenal beliefs, phenomenal beliefs are not knowledge; so

(3) If phenomenal realism is true, phenomenal beliefs are not knowledge.

Some phenomenal realists might deny the first premise: a type-B materialist could hold that experiences have effects on beliefs by virtue of their identity with physical states, and a property dualist could hold that these effects proceed through a fundamental causal connection between the phenomenal and physical domains, or through a fundamental causal connection among nonphysical mental states. But for the purposes of the argument, I will assume the version of phenomenal realism that makes answering the argument as hard as possible, so I will rule out these responses. In particular, I will assume epiphenomenalism, according to which the phenomenal has no effects on the physical domain.[*]

*[[I am not endorsing epiphenomenalism, but I regard it as one of the three serious options that remain once one accepts phenomenal realism and rules out type-B materialism and idealism. The other two are interactionism and a Russellian "panprotopsychism". See Chalmers 2002a.]]

The view I have outlined makes it easy to see why this argument fails, even against an epiphenomenalist. Whatever the status of the first premise, the second premise is false. The second premise assumes that a causal connection between experience and phenomenal belief is required for the latter to count as knowledge. But if what I have said is correct, the connection between experience and phenomenal belief is tighter than any causal connection: it is constitution. And if a causal connection can underwrite knowledge, a constitutive connection can certainly underwrite knowledge too.

Even without appealing to constitution, the epiphenomenalist can respond reasonably to this argument by appealing to the notion of acquaintance, and arguing that a subject's acquaintance with experience can noncausally justify a phenomenal belief. (I used this strategy in *The Conscious Mind*.) But when the role of constitution is made clear, the reply becomes even stronger. Acquaintance and constitution together enable a theoretical model of the justification of phenomenal belief (as above), a model that is compatible with epiphenomenalism. And any residual worries about the lack of an appropriate connection between the experience and the belief are removed by the presence of a constitutive connection.

This first argument is only a subsidiary argument in Shoemaker's discussion. Shoemaker's main argument specifically concerns the possibility of absent qualia. His argument involves functional duplicates and conceptual possibility, but as before I will modify these details to involve physical duplicates and metaphysical possibility, thus making the argument harder to answer. The modified argument runs roughly as follows:

(1) If phenomenal realism is true, then every conscious being has a possible zombie twin.

(2) If zombies are possible, they have the same phenomenal beliefs as their conscious twins, formed by the same mechanisms.

(3) If zombies are possible, their phenomenal beliefs are false and unjustified.

(4) If it is possible that there are beings with the same phenomenal beliefs as a conscious being, formed by the same mechanism, where those phenomenal beliefs are false and unjustified, then the conscious being's phenomenal beliefs are unjustified; so

(5) If phenomenal realism is true, every conscious being's phenomenal beliefs are unjustified.

Some phenomenal realists could respond by denying premise 1 and holding that zombies are impossible. But even the conceptual possibility of functional duplicates with absent qualia is arguably enough to make an analogous argument go through, if there are no other problems. Premise 3 is relatively unproblematic. Perhaps one could argue that a zombie's phenomenal beliefs have some sort of justification, but the conclusion that our phenomenal beliefs are no more justified than a zombie's would be strong enough for an opponent. Disputing premise 4 holds more promise. If one accepts an acquaintance model of justification, one might hold that the justification of a phenomenal belief does not supervene on its mechanism of formation. (I used this strategy in *The Conscious Mind*.) But given what has gone before in this paper, by far the most obvious reply is to dispute premise 2. There is no reason to accept that zombies have the same phenomenal beliefs as their conscious twins, and every reason to believe that they do not.

It is by no means obvious that zombies have beliefs at all. The basis of intentionality is poorly understood, and one might plausibly hold that a capacity for consciousness is required for intentional states. But even if we allow that zombies have beliefs, it is clear that a zombie cannot share a conscious being's phenomenal beliefs. The content of a conscious being's direct phenomenal beliefs is partly constituted by underlying phenomenal qualities. A zombie lacks those qualities, so it cannot have a phenomenal belief with the same content.

Let us take the case of Zombie Mary, where we recombine thought-experiments in the obvious way. Assuming that Zombie Mary has a belief where Mary has a direct phenomenal belief, what sort of content does it have? Mary has a belief with the content $E=R$, and Inverted Mary has a belief with the content $E=G$. Let us focus on the direct phenomenal concepts R and G , and their zombie counterpart. It is obvious that Zombie Mary's concept is neither R nor G : if it has content at all, it has a different content entirely. I think that the most plausible view is that the zombie's concept is *empty*: it has no content. On the view I have been outlining, a phenomenal quality can be thought of as filling a slot that is left open in the content of a direct phenomenal concept, and thus contributing its content. If there is no phenomenal quality to fill the slot, as in Zombie Mary's case, the concept will have no content at all.

What about Zombie Mary's analog of Mary's direct phenomenal belief $E=R$? It is not obvious that a

zombie can possess a demonstrative phenomenal concept: a concept whose content is that of 'this experience' seems to require a concept of experience, which a zombie may lack. But even if a zombie could possess a demonstrative phenomenal concept, any such concept would fail to refer (like failed demonstratives in other domains). And more importantly, the other half of the identity (the zombie's analog of *R*) would be empty. So Zombie Mary's belief would be entirely different from Mary's belief.

It is natural to wonder about the truth-value of Zombie Mary's belief. Clearly her belief is not true. I would say that it is either false or empty, depending on one's view about beliefs involving empty concepts. The latter view is perhaps the most plausible, since it seems that Zombie Mary's belief has no propositional content to evaluate. As for Zombie Mary's "new knowledge": it is clear that she gains no propositional knowledge (though she may think that she does). One might see her as in the position that type-A materialists, and in particular proponents of the "ability hypothesis", hold that we are in in the actual world. When Zombie Mary first sees a flower, she may gain certain abilities to recognize and discriminate, although even these abilities will be severely constrained, since they cannot involve experiences.

This is enough to see that the epistemological argument against phenomenal realism does not get off the ground. A zombie clearly does not have the same phenomenal beliefs as its conscious twin in general; and its corresponding beliefs are not even formed by the same mechanism, since constitution by a phenomenal quality plays a central role in forming a direct phenomenal belief. So the second premise is false, and there is no bar to the justification of direct phenomenal beliefs.[*]

*[[Conee (1985) and Francescotti (1994) also respond to Shoemaker's argument by denying the equivalent of premise 2, although for somewhat different reasons.]]

What about other phenomenal beliefs? We have seen that standing phenomenal concepts differ between twins, and that their content is plausibly constituted either by phenomenal properties or by dispositions involving those properties. A zombie lacks all phenomenal properties, so it is plausible that its analogs of standing phenomenal concepts will be empty, too. So beliefs involving standing phenomenal concepts are also immune from this argument.

What about the standing concept of *experience* (or *qualia* or *phenomenal consciousness*) generally? In this case there is no difference in content between conscious twins. But it remains plausible that phenomenal properties and the capacity to have them play a crucial role in constituting its content, just as they do for specific standing phenomenal concepts. And it is equally plausible that the zombie's analog of this standing concept is empty.[*] So beliefs involving the standing concept of experience (such as *I am conscious*) are equally unthreatened by this argument. The same goes for beliefs involving concepts in which the concept *experience* plays a part, such as relational phenomenal concepts, and perhaps demonstrative phenomenal concepts.

*[[This is relevant to an argument against conceivability arguments for property dualism given by Balog (1999). Balog maintains that a zombie could make a conceivability argument with the same form, with true premises and a false conclusion, so the argument form must be invalid. Balog's argument requires as a premise the claim that a zombie's

assertion 'I am phenomenally conscious' (and the like) expresses a truth. But the discussion here suggests that it is much more plausible that the assertion is false or truth-valueless. This is plausible on independent grounds: in a zombie world, when a zombie realist asserts (an analog of) 'Qualia exist', and a zombie eliminativist asserts 'Qualia do not exist', it seems clear that the zombie eliminativist is closer to being correct. If so, Balog's argument fails.

Balog also discusses "Yogis", creatures that make a form of direct reference to brain states without this being mediated by phenomenology. I think it is clear that Yogis have at most a sort of demonstrative concept (roughly: "*this* inner state"), and do not have the analog of pure phenomenal concepts. For these concepts, no analogous epistemic gap arises. For example, given full physical and indexical information, Yogis will be in a position to know all truths involving the concepts in question.]]

How are these beliefs justified? For beliefs involving standing phenomenal concepts, such as $E=S$, we have seen that one reasonable model involves inference from $E=R$ and $R=S$. Here, the former belief is justified by acquaintance and constitution, and the second belief is justified a priori by virtue of its content. These two beliefs combine by virtue of the common element R to justify the belief $E=S$. (One can also hold that $E=S$ is justified directly by acquaintance, at cost of losing the special contribution of constitution.) One can justify general beliefs of the form E is a phenomenal property in much the same way, given that R is a phenomenal property is a priori.

From here, beliefs such as *I am conscious* are a short leap away. The leap is nontrivial, as there are distinctive problems about the epistemology of the self: witness Hume's skepticism about the self, and Lichtenberg's point that in the cogito, Descartes was entitled only to *There is thought*, not to *I think*. I have nothing special to say about these epistemological problems. But assuming that these problems can be solved, it is not implausible that a belief such as *If E exists, I have E* is justified (perhaps a priori). Then the whole range of first-person phenomenal beliefs lies within reach.

(If one takes direct phenomenal beliefs as truly foundational, one might even suggest that the cogito should have a three-stage structure: from $E=R$ (or some such), to *I have E*, to *I exist!*)

As for beliefs involving relational phenomenal concepts: presumably beliefs such as $S=red_I$, where S is a standing pure concept of phenomenal redness, will be justified a posteriori, perhaps by inference from the observation that the relevant paradigmatic objects typically cause one to experience instances of S . And beliefs of the form $S=red_C$ will be justified at least insofar as $red_I=red_C$ is justified. Of course for the first sort of belief to be justified, skeptical problems about the external world (and about the self) must be overcome, and for the second sort of belief to be justified, skeptical problems about other minds must be overcome. I have nothing special to say about these problems here. But assuming that these problems can be dealt with, then both general relational phenomenal beliefs (e.g. $S=red_C$) and particular relational phenomenal beliefs (e.g. $E=red_C$) will be justified straightforwardly.

It seems, then, that a wide range of phenomenal beliefs can be justified by inference from direct phenomenal beliefs (such as $E=R$), a priori phenomenal beliefs (such as $R=S$ and perhaps *If E exists, I have E*), and a posteriori phenomenal beliefs such as ($S=red_I$ and $S=red_C$). I have given a model for the justification of direct phenomenal beliefs. Phenomenal realism, and even epiphenomenalism, seem to pose no particular problem for the justification of the a priori phenomenal beliefs (or at least no

distinctive problem that does not arise for a priori justification on any view). And the same goes for the justification of the a posteriori phenomenal beliefs. Even if experience plays no causal role, this does not matter. Experiences have no special role in justifying the a priori beliefs, and the justification of the a posteriori beliefs can be seen as derivative on beliefs of the form $E=S$ (which are already accounted for), plus general methods of external observation and inductive inference.

So all we need to justify all these beliefs is the justification of direct phenomenal beliefs, the justification of a priori beliefs in virtue of their content, and the justification involved in inference, observation, and induction. There are no special problems in any of these matters for the phenomenal realist. One might think that inference poses a problem for the epiphenomenalist: how do $E=R$ and $R=S$ justify $E=S$ if the content of R is partly constituted by an epiphenomenal quality, and if inference requires causation? But this is no problem: R acts as a middle term and its content is not required to play any special causal role. We can think of the inference in question as being *E is R , which is S , so E is S* . Here the content of R is inessential to the validity of the inference: as long as the premises are justified, the conclusion will be justified.

Perhaps the only residual epistemological issue concerns the persistence of standing phenomenal concepts. One might worry if S is partly constituted by an element that is epiphenomenal, then even if one acquires a justified belief - say of the form *roses cause S* - at one time, it is not clear how this justification carries over to instances of a belief with that content at a later time. It is plausible that more than a match in content is required for justification: the later belief must be in some sense the "same" belief, or at least a "descendant" belief, involving the "same" (or "descendant") concepts. The same sort of issue arises with inference of the sort in the previous paragraph. Whether or not *E is S* is wholly distinct from the two premises, we certainly want later beliefs of the form *that was S* to be justified, and to play a role in further inferences in turn. But this arguably requires that the later concept be a "descendant" of the earlier concept in a sense that allows beliefs involving the later concept to inherit justification from beliefs involving the earlier concepts.

In response: I have no good account of what it is for one token of a concept to be a "descendant" of another, in a manner that allows it to inherit justification.[*] Nor, I think, does anyone. Clearly more than sameness in content is required: if a new concept with the same content were to be formed *de novo*, no justification would be inherited. So some sort of natural connection between concept tokens is required. But it is plausible that this sort of connection need only require an appropriate causal connection between the physical vehicles of the concept, along with an appropriate match in content: it is not required that the elements constituting the content of the initial concept do any distinctive causal work.

*[[This sort of persistence relation among tokens is central to our use of concepts and beliefs, but has received less discussion than it might have. In effect, it introduces a "typing" of concepts and beliefs that is more fine-grained than a mere typing by content, but less fine-grained than a typing by numerical identity of tokens. This sort of typing was already tacit in my earlier discussion, when I said that direct phenomenal concepts do not persist beyond the lifetime of an experience, but that standing phenomenal concepts do.]]

To see this, consider the persistence of concepts on an externalist view, where content is constituted by

external factors that may lie in the distant past. Here, the factors that constitute the content of two tokens of the concept will play no distinctive role in causally connecting the tokens, since those factors lie in the distant past. The persistence will instead be supported by appropriate connections between the tokens' physical vehicles. It is plausible that the phenomenal realist, and the epiphenomenalist, can say something similar: conceptual persistence is underwritten by natural connections among vehicles, perhaps along with an appropriate match in content. Of course it would be desirable to have a full positive account of this sort of conceptual persistence, but it seems that there is no distinctive problem for the phenomenal realist here.

Further questions concern the justification of beliefs about the representational content of experiences, and the role phenomenal beliefs might play in justifying beliefs about the external world. I will not say anything about these issues here. But it is plausible that these issues pose mere challenges for the phenomenal realist to answer, rather than posing distinctive arguments against it. The distinctive epistemological problems for phenomenal realism have been removed.

4.4 "The Myth of the Given"

A traditional view in epistemology and the philosophy of mind holds that experiences have a special epistemic status that renders them "given" to a subject. This epistemic status is traditionally held to give phenomenal beliefs a special status, and sometimes to allow experiences to act as a foundation for all empirical knowledge. In recent years, this sort of view has often been rejected. The *locus classicus* for this rejection is Wilfrid Sellars' "Empiricism and the Philosophy of Mind" (1956), which criticized such views as involving "The Myth of the Given". Sellars' (deliberately abusive) term for the view has caught on, and today it is not uncommon for this label to be used in criticizing such views as if no further argument is necessary.

I do not know whether my view is one on which experiences are "given". It does not fit Sellars' official characterization of the given (as we will see), and there are other characterizations which it also does not fit. But the term "given" (and in particular "myth of the given") often shifts to encompass many different views, and it may well be that my view shares something of the spirit of the views that were originally criticized under this label. So rather than trying to adjudicate the terminological issue, we can simply ask: are any of the arguments that have been put forward against the "given" good arguments against the view I have put forward here?

Here one runs up against the problem that clear arguments against the "given" are surprisingly hard to find. There are many suggestive ideas in Sellars' paper, but few explicit arguments. When arguments appear, they often take the form of suggesting alternative views, rather than directly criticizing an existing view. But there is at least one clear argument against the "given" in Sellars' paper. This is his famous "inconsistent triad". This was intended as an argument against sense-datum theories, but it clearly applies to a wider class of views.

It is clear from the above analysis, therefore, that classical sense-datum theories [...] are

confronted by an inconsistent triad made up of the following three propositions:

A. *x senses red sense content s entails x non-inferentially knows that s is red.* B. The ability to sense sense contents is unacquired. C. The ability to know facts of the form *x is phi* is acquired.

A and B together entail not-C; B and C entail not-A; A and C entail not-B. (Sellars 1956, section 6)

It is clear how the view I have put forward should deal with this inconsistent triad: by denying A. I have said nothing about just which mental capacities are acquired or unacquired, but on the view I have put forward, it is clearly possible to have experiences without having phenomenal beliefs, and therefore without having knowledge of phenomenal facts. On my view, phenomenal beliefs are formed only rarely, when a subject attends to his or her experiences and makes judgments about them. The rest of the time, the experiences pass unaccompanied by any phenomenal beliefs or phenomenal knowledge.

Underlying Sellars' critique is the idea that knowledge requires concepts, and that experiences do not require concepts, so that having experiences cannot entail having knowledge. The view I have put forward is compatible with all of this. On my view, experiences require little cognitive sophistication, and in particular do not require the possession of concepts. There may be some experiences that require concepts (for example, the experience of a spoon as a spoon), but not all experiences do. No concepts are required to experience phenomenal redness, for example. Knowledge of facts requires belief, however, and belief requires the possession of concepts. So experience does not entail knowledge.

Sellars associated the "given" most strongly with the acceptance of (A), and the denial of (A) is what he argues for himself. In discussing the possibility that a sense-datum theorist might deny (A), all he says is:

He can abandon A, in which case the sensing of sense contents becomes a noncognitive fact - a noncognitive fact, to be sure which may be a necessary condition, even a logically necessary condition, of non-inferential knowledge, but a fact, nevertheless, which cannot constitute this knowledge.

On my view, all this is correct. Experiences do not, on their own, constitute knowledge. They play a role in *justifying* knowledge, and they play a role in *partly* constituting the beliefs that qualify as knowledge, in combination with other cognitive elements. But experiences themselves are to be sharply separated from beliefs and from items of knowledge. So none of this provides any argument against my view.

(On my reading, a number of the sense-data theorists also deny (A), making clear distinctions between the sort of nonconceptual epistemic relation that one stands in by virtue of having an experience and the sort of conceptual epistemic relation that one has when one knows facts. Such theorists clearly avoid the conflation between experience and knowledge that Sellars accuses sense-datum theorists of.)

Curiously, Sellars never discusses the possibility that experiences could justify knowledge without entailing knowledge. It seems clear that he would reject such a view, perhaps because he holds that only conceptual states can enter into justification, but this is never made explicit in his article.[*]

*[[The one further part of Sellars' article that may be relevant to the view I have put forward is part VI (sections 26-29), where he addresses the traditional empiricist idea that experience involves awareness of determinate repeatables. This is closely related to my claim that experience involves acquaintance with properties. Sellars does not provide any direct argument against this view, however. He simply notes (sections 26-28) that Locke, Berkeley, and Hume take this thesis as a presupposition rather than a conclusion (they use it to give an account of how we can be aware of determinable repeatables). And then he asserts (section 29) that this awareness must either be mediated by concepts (e.g. through the belief that certain experiences resemble each other, or that they are red) or be a purely linguistic matter. He gives no argument for this claim, which I think should be rejected. On my view, our acquaintance with qualities requires neither concepts nor language.]]

Although Sellars does not argue explicitly against this sort of view, such arguments have been given by a number of later philosophers writing in the same tradition. In particular, there is a popular argument against any view on which experiences are nonconceptual states that play a role in justifying beliefs. This argument, which we might call the *justification dilemma*, has been put forward by Bonjour (1969), Davidson (19xx), and McDowell (1994), among others. We can represent it as follows:

- (1) There can be no *inferential* relation between a nonconceptual experience and a belief, as inference requires connections within the conceptual domain.
- (2) But a mere *causal* relation between experience and belief cannot justify the belief; so
- (3) Nonconceptual experiences cannot justify beliefs.

The first premise is plausible, as it is plausible that inference is mediated by concepts. The status of the second premise is much less clear. While it is plausible that the mere existence of a causal connection does not suffice to justify a belief, it is far from clear that the right *sort* of causal connection could not serve to justify a belief. McDowell says that a causal connection "offers exculpation where we wanted justification". But clearly causal connections cannot involve mere exculpation simply by virtue of being causal connections, as the case of inference shows: here a causal connection of the right kind between states can be seen to justify. So further argument is required to show that no other sort of causal connection (perhaps with subtle constraints on the content of a belief and on the relationship between belief and experience) can provide justification.

But in any case, even if the two premises are accepted, the conclusion does not follow. An option has been missed: inference and causation do not exhaust the possible justifying relations between nonconceptual experiences and beliefs. On my view, the relation in question is not inference or causation, and neither is it identity or entailment, as on the views that Sellars criticized. Rather, the relation is partial constitution.

I have already given a model of how the justification of a direct phenomenal belief by an experience works, involving three central elements that parallel the three central elements in the case of inference. The analog of the causal element is the constitutive connection between experience and belief; the analog of the content element is the match between epistemic content of belief and quality of experience; and the analog of the epistemic element is the subject's acquaintance with the phenomenal quality. If the model of justification by inference is accepted, there is no clear reason why this model should be rejected.

Some philosophers hold that only a conceptual state can justify another conceptual state. But as with the thesis that only a belief can justify another belief, it is not clear why this thesis should be accepted. It is not supported pretheoretically: pretheoretically, there is every reason to hold that experiences are nonconceptual and can justify beliefs. And there is no clear theoretical support for this claim, either. Proponents sometimes talk of "the space of reasons" in this context, but the slogan alone does not convert easily into an argument. McDowell suggests that justifications for our beliefs should be *articulable*, which requires concepts; but as Peacocke (2001) points out, we can articulate a justification by referring to a justifying experience under a concept, whether or not the experience itself involves concepts. Perhaps the central motivation for the thesis lies in the fact that we have a clear theoretical model for conceptual justification, but not for other sorts of justification. But again, this is a weak argument, and again, the exhibition of a theoretical model ought to remove this sort of worry.

In any case: the view I have put forward avoids Sellars' central version of the given (an entailment from experience to knowledge), and Bonjour/Davidson/McDowell's central version of the given (a mere causal connection), along with the arguments against those views. It may be that the view I have put forward accepts a "given" in some expanded sense. But the substantive question remains: are there good arguments against the given that are good arguments against this view? I have not been able to find such arguments, but I would welcome hearing candidates.

5 Further questions

I have drawn a number of conclusions about the content and epistemology of phenomenal beliefs. It is natural to ask whether these conclusions apply more generally.

First, regarding content: I have argued that the content of phenomenal concepts and phenomenal beliefs is conceptually irreducible to the physical and functional, because this content itself depends on the constitutive role of experience. Does this sort of irreducibility extend to other concepts or beliefs? Is the content of concepts and beliefs irreducible to the physical and functional quite generally?

There is one class of concepts for which such a conclusion clearly follows. This is the class of concepts that have phenomenal concepts as constituents. Such concepts might include *the tallest conscious being in this room*, *the physical basis of consciousness*, and *the external cause of R*, where *R* is a pure concept of phenomenal redness. More generally, insofar as a concept has conceptual ties with phenomenal concepts, so that claims involving that concept conceptually and nontrivially entail claims involving pure phenomenal concepts, then the content of such a concept will be irreducible in a similar way.

It is arguable that many or most of our perceptual concepts have this feature. At least some concepts of external colors can be analyzed roughly as *the property causally responsible for C in me*, where *C* is a pure concept of a phenomenal color. Things are more complex for community-level concepts. Here it is more plausible that an external color concept might be analyzed in terms of community-wide relations to a nonspecific phenomenal concept: perhaps *the property causally responsible for the dominant sort of visual experience caused by certain paradigmatic objects in this community*, or something like that. But this still has the concept of visual experience as a constituent, and so will still have functionally irreducible content. The alternative is that external color concepts might be analyzed in terms of their relations to certain *judgments* or other nonexperiential responses, in which case the reducibility or irreducibility will not be so clear. I will not adjudicate this matter here, but my own view is that while there may be some perceptual concepts without an obvious phenomenal component, many or most of the perceptual concepts that we actually possess have such a component.

One might try to extend this further. In the case of theoretical concepts from science, for example, one can argue that these have conceptual ties to various perceptual concepts (as the Ramsey-Lewis analysis of theoretical concepts suggests). If so, and if the perceptual concepts in question have irreducible content, it is arguable that these concepts have irreducible content. And one might argue for conceptual ties between intentional concepts and phenomenal concepts, and between social concepts and intentional concepts, so that a wide range of social concepts will turn out to have irreducible content. If this is right, then a being without consciousness could have at best impoverished versions of these concepts, and perhaps no such concepts at all.

This sort of argument will not work for all concepts. Many mathematical or philosophical concepts have no obvious tie to phenomenal concepts, for example. And in fact there is good reason to think that some concepts do not have an phenomenal component. If all concepts have a phenomenal component, it would be hard to avoid the conclusion that all concepts are *entirely* constituted by phenomenal concepts, which would lead naturally to phenomenalism or idealism. My own view is that certain central concepts, such as that of causation, have no deep phenomenal component at all. Once this is recognized, it becomes clear that even if a wide range of concepts have a phenomenal component, only a small number of them are entirely phenomenal.

Even if some concepts have no phenomenal component, it is not out of the question that their content might still be irreducible. One intriguing possibility is that something about a subject's phenomenal states could be central to a subject's *possessing* a concept such as that of causation, or certain mathematical concepts, even though these concepts do not refer to phenomenal states as part of their content. (Compare a reductive view on which neural states might constitute the content of concepts that do not refer to neural states.) There is at least some intuition that a capacity for consciousness may be required to have concepts in the first place; and it is not obviously false that phenomenology plays a role in the possession of even nonphenomenal concepts.

Such a thesis would require much further argument, of course, and I am not certain whether it is true. But even if it is false, the more limited thesis that phenomenology plays a role in constituting the content of

phenomenal concepts, and that phenomenal concepts play a role in determining the content of a wide range of other concepts, has significant consequences. If even the more limited thesis is true, then the project of giving a functional analysis of intentionality cannot succeed across the board, and a central role must be given to phenomenology in the analysis of intentional content.

Second, epistemology: I have in effect argued for a sort of limited foundationalism within the phenomenal domain. Direct phenomenal beliefs are in a certain sense foundational: they receive justification directly from experience, and their *prima facie* justification does not rely on other beliefs. And I have argued that direct phenomenal beliefs can justify at least some other phenomenal beliefs in turn, when aided by various sorts of *a priori* reasoning. Does this give any support to foundationalism about a broader class of empirical beliefs, or about empirical knowledge in general?

Nothing I have said implies this. This gap between phenomenal knowledge and knowledge of the external world remains as wide as ever, and I have done nothing to close it. The framework here is compatible with various standard suggestions: that phenomenology might justify external beliefs through inference to the best explanation, or through a principle that gives *prima facie* justification to a belief that endorses an experience's representational content. But so far, the framework outlined here does nothing special to support these suggestions or to answer skeptical objections. And the framework is equally compatible with many alternative nonfoundationalist accounts of our knowledge of the external world.[*]

*[[A particular problem in extending this account to a general foundationalism is that we do not usually form direct phenomenal beliefs associated with a given experience, so such beliefs are not available to help in justifying perceptual beliefs. (Thanks to Alvin Goldman for discussion on this point.) Here there are a few alternatives: (1) deny that perceptual beliefs are usually justified in the strongest sense, but hold that such justification is available; (2) hold that the mere availability of justifying direct phenomenal beliefs confers a sort of justification on perceptual beliefs; or (3) extend the account so that perceptual experiences can justify perceptual beliefs directly, through a constitutive connection to perceptual concepts analogous to the connection to phenomenal concepts. I explore the third possibility in forthcoming work on the content of perceptual experience and perceptual belief.]]

Still, this framework may help to overcome what is sometimes taken to be the largest problem for foundationalism: bridging the gap between experience and belief. I have argued that an independently motivated account of the role of experience in phenomenal belief, and of subject's epistemic relations to them, has the resources to solve this problem, by exploiting the paired notions of constitution and acquaintance.

Any plausible epistemological view must find a central role for experience in the justification of both beliefs about experience and beliefs about the world. If what I have said in this paper is correct, then we can at least see how experience gains a foothold in this epistemic network. Many other problems remain, especially regarding the relationship between experience and beliefs about the external world. But here, as in the case of phenomenal belief, a better understanding of the relationship between experience and belief may take us a long way.

Acknowledgements

I have given versions of this paper (starting in 1997) at Antwerp, ANU, Arizona, Delaware, Fribourg, Miami, Munich, Princeton, Sydney, UC Santa Cruz, the APA (Pacific Division), Metaphysical Mayhem (Syracuse), and the World Congress in Philosophy. Thanks to many people in audiences there, and elsewhere. Special thanks to Mark Johnston, Martine Nida-Rumelin and Daniel Stoljar for lengthy discussions.

References

- Austin, D.F. 1990. *What's the Meaning of "This"?* Ithaca, NY: Cornell University Press.
- Balog, K. 1999. Conceivability, possibility, and the mind-body problem. *Philosophical Review*.
- Bayne, T. 2001. Chalmers, the acquaintance relation, and phenomenal judgment. *Philosophy and Phenomenological Research* 62:407-19.
- BonJour, L. 1978. Does empirical knowledge have a foundation?
- BonJour, L. 1969. *Knowledge, Justification, and Truth: A Sellarsian Approach to Epistemology*. Ph.D. Dissertation, Princeton University.
- Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D.J. 2002a. Consciousness and its place in nature. In (S. Stich & F. Warfield, eds) *The Blackwell Guide to the Philosophy of Mind*. Blackwell. [consc.net/papers/nature.html]
- Chalmers, D.J. 2002b. Does conceivability entail possibility? In (T. Gendler & J. Hawthorne, eds) *Conceivability and Possibility*. Oxford University Press. [consc.net/papers/conceivability.html]
- Chalmers, D.J. 2002c. The components of content. In (D. Chalmers, ed) *The Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press. [consc.net/papers/content.html]
- Chalmers, D.J. (forthcoming). The nature of epistemic space. [consc.net/papers/espace.html]
- Chisholm, R. 1957. *Perceiving: A Philosophical Study*.
- Conee, E. 1985. The possibility of absent qualia. *Philosophical Review* 94:345-66.
- Francescotti, R.M. 1994. Qualitative beliefs, wide content, and wide behavior. *Nous* 28:396-404.
- Fumerton, R. 1995. *Metaepistemology and Skepticism*. Rowman and Littlefield.

- Gertler, B. (forthcoming). Introspecting phenomenal states. *Philosophy and Phenomenological Research*.
- Hawthorne, J. (forthcoming). Advice to physicalists. *Philosophical Studies*.
- Ismael, J. 1999. Science and the phenomenal. *Philosophy of Science*.
- Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136.
- Kaplan, D. 1989. Demonstratives. In (J. Almog, J. Perry, and H. Wettstein, ed.) *Themes from Kaplan*. New York: Oxford University Press.
- Kripke, S.R. 1981. *Wittgenstein on Rules and Private Language*. Harvard University Press.
- Loar, B. 1997. Phenomenal states (second version). In (N. Block, O. Flanagan, & G. Güzelidere, eds) *The Nature of Consciousness*. MIT Press.
- McDowell, J. 1995. *Mind and World*. Harvard University Press.
- Nida-Rumelin, M. 1995. What Mary couldn't know: Belief about phenomenal states. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schöningh.
- Nida-Rumelin, M. 1996. Pseudonormal vision: An actual case of qualia inversion? *Philosophical Studies* 82:145-57.
- Nida-Rumelin, M. 1997. On belief about experiences: An epistemological distinction applied to the knowledge argument. *Philosophy and Phenomenological Research*.
- Peacocke, C. 2001. Does perception have a nonconceptual content? *Journal of Philosophy* 98:239-64.
- Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. MIT Press.
- Pollock, J. 1986. *Contemporary Theories of Knowledge*. Rowman and Littlefield.
- Raffman, D. 1995. On the persistence of phenomenology. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schöningh.
- Russell, B. 1910. Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society* 11:108-128.
- Sellars, W. 1956. Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of*

Science 1:253-329. Reprinted as *Empiricism and the Philosophy of Mind*. Harvard University Press, 1997.

Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:291-315.

Wittgenstein, L. 1953. *Philosophical Investigations*.

Appendix

What follows is a brief and simplified introduction to the two-dimensional semantic framework as I understand it. See also Chalmers (2002c; forthcoming).

Let us say that it is epistemically possible in the broad sense that S if the hypothesis that S is not ruled out a priori. Then there will be a wide space of epistemic possible hypotheses (in the broad sense; I omit the qualifier in what follows). Some of these will conflict with each other; some of them will be compatible with each other; and some will subsume each other. We have a systematic way of evaluating and describing epistemic possibilities that differs from our way of evaluating and describing subjunctive counterfactual possibilities. It is this sort of evaluation and description that is captured by the first dimension of the two-dimensional framework.

It is epistemically possible that water is not H₂O, in the broad sense that this is not ruled out a priori. And there are many specific versions of this epistemic possibility: intuitively, specific ways our world could turn out such that if they turn out that way, it will turn out that water is not H₂O. Take the XYZ-world, one containing superficially identical XYZ in place of H₂O. It is epistemically possible that our world is the XYZ-world. When we consider this epistemic possibility - that is, when we consider the hypothesis that *our* world contains XYZ in the oceans, and so on - then this epistemic possibility can be seen as an instance of the epistemic possibility that water is not H₂O. We can rationally say "if our world turns out to have XYZ in the oceans (etc.), it will turn out that water is not H₂O". The hypothesis that the XYZ-world is actual rationally entails the belief that water is not H₂O, and is rationally inconsistent with the belief that water is H₂O.

Here, as with subjunctive counterfactual evaluation, we are considering and describing a world, but we are considering and describing it in a different way. In the epistemic case, we consider a world *as actual*: that is, we consider the hypothesis that our world is that world. In the subjunctive case, we consider a world *as counterfactual*: that is, we consider it as a way things might have been, but (probably) are not. These two modes of consideration of a world yield two ways in which a world might be seen to make a sentence or a belief true. When the XYZ-world is considered as actual, it makes true 'water is XYZ'; when it is considered as counterfactual, it does not.

In considering a world as actual, we ask ourselves: what if the actual world is really that way? In the broad sense, it is *epistemically* possible that Hesperus is not Phosphorus. This is mirrored by the fact that there are specific epistemic possibilities (not ruled out a priori) in which the heavenly bodies visible in

the morning and evening are distinct; and upon consideration, such epistemic possibilities are revealed as instances of the epistemic possibility that Hesperus is not Phosphorus.

When we consider worlds as counterfactual, we consider and evaluate them in the way that we consider and evaluate subjunctive counterfactual possibilities. That is, we acknowledge that the character of the actual world is fixed, and say to ourselves: what if the world *had been* such-and-such a way? When we consider the counterfactual hypothesis that the morning star might have been distinct from the evening star, we conclude not that Hesperus would not have been Phosphorus, but rather that at least one of the objects is distinct from both Hesperus and Phosphorus (at least if we take for granted the actual-world knowledge that Hesperus is Phosphorus, and if we accept Kripke's intuitions).

Given a statement *S* and a world *W*, the *epistemic intension* of *S* returns the truth-value of *S* in *W* considered as actual. (Test: if *W* actually obtains, is *S* the case?) The *subjunctive intension* of *S* returns the truth-value of *S* in *W* considered as counterfactual. (Test: if *W* had obtained, would *S* would have been the case?) We can then say that *S* is *primarily possible* (or 1-possible) if its epistemic intension is true in some world (i.e. if it is true in some world considered as actual), and that *S* is *secondarily possible* (or 2-possible) if its subjunctive intension is true in some world (i.e. if it is true in some world considered as counterfactual). Primary and secondary necessity can be defined analogously.

For a world to be considered as actual, it must be a *centered* world -- a world marked with a specified individual and time - as an epistemic possibility is not complete until one's "viewpoint" is specified. So a epistemic intension should be seen as a function from centered world to truth-values. For example, the epistemic intension of 'I' picks out the individual at the center of a centered world; and the epistemic intension of 'water' picks out, very roughly, the clear drinkable (etc.) liquid in the vicinity of the center. No such marking of a center is required for considering a world as counterfactual, or for evaluating subjunctive intensions.

Epistemic and subjunctive intensions can be associated with statements in language, as above, and equally with singular terms and property terms. The intension of a statement will be a function from worlds to truth-values; the intension of a term will be a function from worlds to individuals or properties within those worlds. (In some cases, intensions are best associated with linguistic tokens rather than types.)

Epistemic intensions can also be associated in much the same way with the (token) concepts and thoughts of a thinker, all of which can be used to describe and evaluate epistemic possibilities as well as subjunctive counterfactual possibilities. In "The Components of Content" I argue that the epistemic intension of a concept or a thought can be seen as its "epistemic content" (a sort of internal, cognitive content), and that the subjunctive intension captures much of what is often called "wide content".

A crucial property of epistemic content is that it reflects the rational relations between thoughts. In particular, if a belief *A* entails a belief *B* by a priori reasoning, then it will be epistemically impossible (in the broad sense) for *A* to be true without *B* being true, so the epistemic intension of *A* entails the

epistemic intension of B . Further, if an identity $a=b$ is a posteriori for a subject, then it is epistemically possible for the subject that the identity is false, and there will be an epistemic possibility in which the referents of the two concepts involved differ, so the subject's concepts a and b will have distinct epistemic intensions. This applies even to beliefs expressed by a posteriori necessities such as 'water is H₂O' and 'Hesperus is Phosphorus': the epistemic intensions of these beliefs are false at some worlds, so the concepts involved have different epistemic intensions. So epistemic intensions behave something like Fregean senses, individuating concepts according to cognitive significance at least up to the level of a priori equivalence.

(A complication here: on some philosophical views, there may be "strong necessities" whose epistemic intension is false at no world. An example might be 'A god exists', on a theist view on which a god exists necessarily but not a priori. These necessities go well beyond Kripkean a posteriori necessities, and I have argued elsewhere (Chalmers 2002b) that there are no such necessities. If they exist, however, the present framework can accommodate them by moving to a broader class of conceptual or epistemic possibilities, which need not correspond to metaphysical possibilities (see Chalmers (forthcoming) for more details. In the case above, for example, there will be at least a conceptually possible world (or "scenario") in which there is no god; and more generally, any a posteriori belief will have an epistemic intension that is false at some such world.)

Insofar as the two-dimensional framework is used in this paper, it is being applied rather than being discussed or justified in its own right. The discussion here indicates important distinctions among phenomenal concepts whose analysis requires the idea of epistemic content. And importantly, there are epistemological distinctions that turn on these distinctions in content. This reflects a more general phenomenon: the sort of possibility that is most crucial in epistemology is epistemic possibility, and the sort of content that is correspondingly most crucial is epistemic content.

Philosophy 596B: Metaphysics

Mind and Modality

David Chalmers

E-mail: chalmers@arizona.edu

Web: <http://www.u.arizona.edu/~chalmers/>

Overview

When one thinks about the mind-body problem, or about metaphysics more generally, one is quickly led into thinking about possibility and necessity. Materialism, for example, is naturally understood as making a modal claim -- the physical facts necessitate all the facts -- and questions about the truth of materialism turn on deep questions about the nature of possibility and necessity. How extensive is the space of possible worlds? What is the relationship between conceivability and possibility? How can we have cognitive access to matters of possibility and necessity? In this course, we will use modal considerations in the philosophy of mind as a route into the thickets of modality.

We will start by considering some modal arguments against materialism, and by looking at the varieties of possibility and necessity. Then we will concentrate on the phenomenon of a posteriori necessity, and try to understand its roots (with a detour into two-dimensional semantics, and into the logic of conditionals), and we will try to come to grips with the philosophical concept of a possible world. In the second half of the course, we will concentrate on the relationship between conceivability and possibility, and more generally on the relationship between the epistemology and the metaphysics of modality; and we will look into implications for the philosophy of mind and for metaphysics more generally. We will read recent works by a number of philosophers on these issues, within metaphysics, the philosophy of language, and the philosophy of mind. Along the way, I will present portions of some work in progress on the general theme of "modal rationalism".

Readings

- Saul Kripke, *Naming and Necessity*.
- David Chalmers, *The Conscious Mind*. (esp Chap. 1-4)
- David Lewis, *On the Plurality of Worlds*. (esp Chap. 1-3)
- A reading packet with numerous articles.
- A number of articles on the web.

The Lewis book is out of print, unfortunately, so I will include some of it in the reading packet (which won't be needed for a few weeks; I'll let you know about availability as the time approaches). Where

articles are available on the web, I've indicated this below. All the web articles should be accessible from my page of [online papers on consciousness](#) (especially the section on [materialism and modality](#)).

You may also want to consult my [bibliography of contemporary philosophy of mind](#) (especially [section 1.3](#)).

Much of the course will loosely follow the structure of my Princeton lectures on [Mind and Modality](#) (especially the first two lectures),

Mailing list

I will set up a mailing list, modality@listserv.arizona.edu, for class discussion. This will be the main locus of the course while I am in Australia, and will continue to play a role when I return. Everyone will be expected to make reasonably regular contributions to this list (at least one reasonably substantial posting every week), discussing issues arising from the readings, from class discussion, and from the mailing list itself. Discussion from the mailing list is retrospectively included with the associated weeks below.

Schedule

Here is a very approximate week-by-week plan for the course. The correspondence between topics here and weeks will probably be loose. In the first few weeks in particular, things will be done somewhat differently to accommodate the electronic format.

1. Metaphysics and Modality

Varieties of possibility and necessity. Conceptual, metaphysical, natural possibility; a priority; analyticity. The importance of modality in metaphysics. A sampling of modal arguments in metaphysics (causation, external reality, morality, intentionality, consciousness). Modal definitions of materialism.

- TCM, Chapter 2.
- Lewis, *On the Plurality of Worlds*, Chapter 1.

[Week 1 discussion](#)

2. Epistemic Arguments Against Materialism.

Conceivability arguments: the possibility of disembodiment (Descartes), of zombies, of inverted spectra. Epistemic arguments: Jackson's knowledge argument, etc. Conceptual analysis of phenomenal concepts. Objections. Type-A and type-B materialism.

- TCM, Chapter 3.
- Jackson, [Epiphenomenal qualia](#)

[Week 2 discussion](#)

3. Naming and Necessity

Kripke's *Naming and Necessity*. The modal argument for rigid designation and a posteriori necessities. The epistemic argument against descriptive theories of reference. Names and natural kind terms. Application to metaphysical issues.

- Kripke, *Naming and Necessity*.

[Week 3 discussion](#)

4. The 2-D account.

The two-dimensional account of a posteriori necessity. Primary and secondary intensions. Two notions of necessity. Link to apriority. Translation of Kripke's main theses into two-dimensional terms. Related ideas (Kaplan, Stalnaker, etc).

- TCM, Section 2.4
- Chalmers, [The components of content](#)
- Evans, Reference and Contingency
- Davies and Humberstone, Two Notions of Necessity

[Week 4 discussion](#)

5. The Tyranny of the Subjunctive.

The different behavior of subjunctive and indicative conditionals in the analysis of possibilities. Links to two-dimensional semantics. The implicit dominance of the subjunctive in contemporary analyses of necessity. How philosophy would look with indicatives instead. The analysis of subjunctives and indicatives. Epistemic vs. subjunctive possibility.

[Week 5 discussion](#)

6. Applying the 2-D account to the anti-materialist argument.

Type-B materialism: relying on a posteriori necessity to save the materialist. Application of the two-

dimensional analysis of a posteriori necessity to this position. From an epistemic gap to an ontological gap? Loopholes.

- TCM, Chapter 4.
- Chalmers, [Mind and modality, part 1](#)
- Jackson, Finding the Mind in the Natural World.

[Week 6 discussion](#)

7. Materialism and the Metaphysics of Modality (Strong Necessities).

The type-B materialist's escape route: strong necessities. Are there strong necessities? What do they involve? Possible examples and explanations.

- PPR symposium on *The Conscious Mind* (Hill & McLaughlin, Loar, Shoemaker, [Yablo](#)).
- Yablo, [Textbook Kripkeanism and the open texture of language](#)
- Loar, [Phenomenal states](#)
- Hill, Conceivability, Possibility, and the Mind-Body Problem
- Chalmers, [Materialism and the metaphysics of modality](#)

[Week 7 discussion](#)

8. Does Conceivability Entail Possibility?

Varieties of conceivability: prima facie vs. ideal, primary vs. secondary, positive vs. negative, etc. Possible counterexamples to the conceivability/possibility link.

- Chalmers, [Mind and modality, part 2](#)
- Chalmers, [Does conceivability entail possibility?](#)
- Yablo, Is conceivability a guide to possibility?
- van Cleve, Conceivability and the Cartesian argument for dualism

[Week 8 discussion](#)

9. Modal Rationalism and the Scrutability of Truth

Varieties of modal rationalism. Three loopholes: strong necessities, inscrutable truths, open inconceivabilities. The analysis of inscrutability. Links to inscrutability of reference (Quine, Putnam, Benacerraf). Possible examples: inscrutable mathematical truths, epistemic theory of vagueness, etc. Arguments against inscrutability?

[Week 9 discussion](#)

10. Cosmic Hermeneutics

Are all truths entailed a priori by the microphysical (plus phenomenal) truths? Link to inscrutability and modal rationalism. Arguments pro and con.

- TCM, Section 2.5
- Horgan, Supervenience and cosmic hermeneutics
- Block and Stalnaker, [Conceptual analysis and the explanatory gap](#)
- Byrne, [Cosmic hermeneutics](#)

[Week 10 discussion](#)

11. Constructing Possible Worlds

What is a possible world? Concrete vs. "ersatz" conceptions of possible worlds. Constructing "ersatz" worlds. Constructing conceivable worlds, handling epistemic vs. subjunctive possibilities and two-dimensional semantics, etc.

- Lewis, *On the Plurality of Worlds*, Chapters 2 & 3.
- Stalnaker, Possible worlds.
- Adams, Theories of actuality.

[Week 11 discussion](#)

12. Modal Monism and Modal Dualism

The argument for modal rationalism. Do strong necessities commit one to modal dualism? Should there be just one modal primitive? Constitutive links between modal and rational notions. Handling tricky cases, e.g. haecceities.

- Chalmers, [Mind and modality, part 2](#)
- Chalmers, [Materialism and the metaphysics of modality](#)
- Salmon, The logic of what might have been

[Week 12 discussion](#)

13. The Components of Content

Wide content and narrow content. Using two-dimensional semantics to give an account of narrow content. Application to the explanation of behavior, to the semantics of belief ascription, and to belief puzzles. Objections and subtleties.

- Chalmers, [The components of content](#)
- Stalnaker, Narrow Content
- Block, What Narrow Content is Not.

[Week 13 discussion](#)

14. Further Directions

Hyperintensionality and the analysis of epistemic possibility. The theory of concepts, meaning, and understanding. Realism and anti-realism.

15. From Mind to Modality and Back Again

Where are we left in the philosophy of mind? The menu of options re the metaphysics of consciousness. The analysis of mental content. Links between consciousness and content?

Assessment

The major item for assessment will be a paper due at the end of term. There may also be some smaller written assignments along the way, and student presentations in class. All students are expected to participate in class discussion (both online and in person).

Mailing list

I will set up a mailing list, modality@listserv.arizona.edu, for class discussion. Everyone will be expected to make reasonably regular contributions to this list (at least one reasonably substantial posting every week), discussing issues arising from the readings, from class discussion, and from the mailing list itself.

Responses to articles on my work

This page is mostly for brief responses to published or forthcoming articles that discuss my work. I've starred [*] the more technical entries so nonphilosophers can skip them. (The [first entry](#) is probably the best place for nonphilosophers to start.) These responses are "unofficial", but if you'd like to use them in an article, feel free to ask. See also [online discussions of my work](#) for some further papers and other discussion (without responses).

Table of contents: what follows includes three collections of articles replied to elsewhere, followed by articles by [Katalin Balog](#), [Tim Bayne](#), [Ned Block & Robert Stalnaker](#), [Anthony Brueckner](#), [Alex Byrne](#), [Alex Byrne & Ned Hall](#), [Paul Churchland](#), [William Greenberg](#), [Christopher Hill](#), [Jenann Ismael](#), [Mark Johnston](#), [Robert Kirk](#), [Noa Latham](#), [Joseph Levine](#), [Harry Lewis](#), [Andrew Melnyk](#), [Nigel Thomas](#), [John Perry](#), [Robert Stalnaker](#), [Robert Stalnaker](#), [Bram van Heuveln et al](#), and [Stephen Yablo](#). (These links are within this page.)

Journal of Consciousness Studies symposium on the "hard problem". (Articles by [Baars](#), Bilodeau, [Churchland](#), [Clark](#), Clarke, [Crick & Koch](#), [Dennett](#), [Hameroff & Penrose](#), Hardcastle, Hodgson, [Hut & Shepard](#), Libet, Lowe, [MacLennan](#), McGinn, Mills, O'Hara & Scutt, Price, Robinson, Rosenberg, Seager, Shear, Stapp, [Velmans](#), Warner.)

See my [Moving Forward on the Problem of Consciousness](#).

**Philosophy and Phenomenological Research* symposium on *The Conscious Mind*. (Articles by Hill & McLaughlin, Loar, Shoemaker, [Yablo](#).)

See my [Materialism and the Metaphysics of Modality](#).

Reviews of *The Conscious Mind*.

See my page of [reviews of *The Conscious Mind*](#). See also [response to Papineau](#) (TLS), [response to Mulhauser](#) (PSYCHE), [response to Searle](#) (NYRB), [second response to Searle](#).

Katalin Balog. Conceivability, possibility, and the mind-body problem. *Philosophical Review* 108:497-528. Online version: [Conceivability arguments or the revenge of the zombies](#).

Balog appeals to zombies in order show that Jackson's and my anti-materialist arguments are self-defeating. She argues that a zombie could make the same arguments, with true premises but a

false conclusion (materialism is true in the zombie world), and concludes that the arguments are invalid. She locates the problem in the inference from conceivability to possibility, or in the thesis that materialism requires *a priori* entailment.

This is an intriguing argument, but I think the problem with it is clear. Balog's parallel argument requires that a zombie's claim "I am conscious" is *true*; otherwise the argument doesn't get off the ground. Balog supports this by suggesting that the zombie's "consciousness" concept will pick out a physical/functional property to which it is causally related. But I think it is much more plausible that the zombie's claim is *false*. The easiest way to see this is to consider an argument in the zombie world, perhaps between Zombie Chalmers and Zombie Dennett. Zombie Chalmers says "Qualia exist", Zombie Dennett says "Qualia do not exist". Balog's analysis implies that in the zombie world, Zombie Chalmers is right. But this seems wrong. Surely in the zombie world, at least, Zombie Dennett is right.

See ["The Content and Epistemology of Phenomenal Belief"](#) for more on this sort of thing.

***Tim Bayne. [Chalmers on the justification of phenomenal judgment](#). *Philosophy and Phenomenological Research* 62:407-19.**

Bayne focuses on the epistemological discussion in Chapter 5 of my book. There I argue that a reliabilist account of phenomenal knowledge can't work, since a reliabilist account can't deliver certain knowledge. When knowledge is grounded only in a reliable connection, we can't be certain that the reliable connection holds, so we can't rule out alternative skeptical scenarios; but we are certain that we are conscious, and can rule out alternative skeptical scenarios. Bayne argues that the same applies to my "acquaintance" account. He notes that I accept in the book that acquaintance alone doesn't suffice for full justification (acquaintance alone is compatible with fallibility, e.g. in cases of inattention, etc). So various further "background conditions" (e.g. concerning attention, etc) are required for full justification. But we can't know for certain that those background conditions hold; so we can't know for certain that we are justified. So by parity of reasoning, the acquaintance model can't deliver certain phenomenal knowledge, either.

I think this argument rests on a natural misreading of my discussion of reliabilism. The argument is *not*: reliabilism can't deliver certainty that we are fully justified, certainty that we are fully justified is required for certainty, so reliabilism can't deliver certainty. That argument would rest on a problematic "CJ" thesis (analogous to the "KJ" thesis that knowledge requires knowledge that one is justified, but with certainty instead of knowledge). Rather, the argument is: when knowledge is grounded only in a reliable connection, we can't be certain that the reliable connection holds, so we can't rule out alternative skeptical scenarios about what's at the other end of the connection, so we can't have certain knowledge.

The first argument (resting on the CJ thesis) just might deliver an analogous argument against the acquaintance model. But I'll respond by denying the CJ thesis (even accepting CJ, one could also

argue that "positive" certainty of justification doesn't require "negative" infallibility). And an argument analogous to the second argument won't work. Even if an acquaintance theorist accepts that we can't be certain that we are justified, that merely shows that one can't rule out skeptical scenarios in which one's belief in the experience is not fully justified; it doesn't show that one can't rule out skeptical scenarios in which the experiences are not present. Presumably an acquaintance theorist can hold that we're certain about qualia but not about such cognitive matters as full justification, so that we can rule out skeptical scenarios without qualia, but not those without full justification. (To resist this would require the dubious CJ claim.) The analogous move is not open to a reliabilist: by its nature, reliabilism can never deliver "knowledge beyond skepticism". Again, see ["The Content and Epistemology of Phenomenal Belief"](#) for more.

***Ned Block and Robert Stalnaker.** [Conceptual analysis, dualism, and the explanatory gap.](#) *Philosophical Review* 108:1-46.

Frank Jackson and I have written a response to this article, forthcoming in the *Philosophical Review*: ["Conceptual Analysis and Reductive Explanation"](#).

***Anthony Brueckner.** Chalmers' conceivability argument for dualism. *Analysis* 61:187-93, 2001.

This paper is a bit puzzling. It starts by saying that others have misunderstood my argument, and then gives a reasonably (though not completely) accurate reconstruction of some aspects of my two-dimensional modal argument. Then in the last section Brueckner says that I don't establish that the conceivability of zombies entails the relevant possibility. But this is a straightforward consequence of the independently argued thesis that conceivability in the relevant sense entails possibility of primary intension (which is all I need). Brueckner doesn't even discuss my positive case for this thesis, so something has gone awry -- I'm not sure what. It may be that the argument in the book was insufficiently formal and explicit; people have misunderstood it more often than I expected. There are more formal versions, as well as defenses of the relevant principles, in ["Mind and Modality"](#) and especially ["Does Conceivability Entail Possibility"](#).

***Alex Byrne.** [Cosmic hermeneutics.](#) *Philosophical Perspectives* 13, 1999.

Byrne argues against the *a priori* entailment of most macro facts by micro facts. Along the way it addresses two of my arguments: the argument from conceivability, in which I argue that a scenario in which the micro facts are the same but the macro facts are different is inconceivable; and one of the arguments from epistemology, in which I argue that a good enough reasoner given only the micro facts would be in a position to ascertain the macro facts. Byrne's main objection to the conceivability argument is that given that we don't yet know just what the micro facts are, we can't know that the scenario in question is inconceivable. His main objection to the epistemological argument is that the strategy I outline relies on the reasoner reaching certain intermediate high-level conclusions, and that the entailment of these conclusions is just as questionable as the entailment of the original macro facts.

I think Byrne misses a key aspect of the arguments in the book (which are admittedly brief). The argument in general is to argue that the micro facts specify micro structure and dynamics, that micro structure and dynamics *a priori* entail macro structure and dynamics, and that macro structure and dynamics entails most macro facts (except those that depend on mentality or indexicals). The middle level of macro structure and dynamics can be thought of as a "geometric" characterization of the structure and evolution of macro objects -- their overall shapes, masses, positions, causal relations, etc. As long as microphysics has information about mass, position, etc, of micro objects, the information about macro objects will be easily derivable from the micro information (e.g. by considerations about macroscopic mass densities in various locations, and so on). So this point is robust over microphysical theories. The second entailment is made plausible not least by the fact that this sort of macro structure and dynamics (plus facts about appearance, which I can also help myself to) is all the information we have to go on in ordinary perception.

Byrne's objection to the conceivability argument can be rebutted by noting that however microphysics turns out, it will be inconceivable that there be the microstructure without the macrostructure, and it will be inconceivable that there be the macrostructure without the macro facts. His objection to the epistemological argument can be rebutted by noting that the reasoner will easily be able to ascertain facts about macrostructure, and from facts about macrostructure (plus facts about appearance) they will be able to ascertain the macro facts in question (just as we can through perceptual information). Byrne does not address my second epistemological argument, which is that the absence of a priori entailment of macro facts will lead to a sort of skeptical problem concerning those macro facts (two epistemic possibilities with the same micro facts and different macro facts would "look" just the same) that we are not in fact faced with in most cases, except in the case of consciousness. All this is now discussed at greater length in ["Conceptual Analysis and Reductive Explanation"](#).

Alex Byrne and Ned Hall. [Chalmers on consciousness and quantum mechanics](#). *Philosophy of Science* 66:370-90, 1999.

This paper addresses the "dessert" in my book, in which I try to use the ideas about consciousness to support a version of the Everett interpretation of quantum mechanics, on which the world is a giant superposition but on which there are minds with discrete experiences like ours. I argue that my principle of organizational invariance predicts that a superposition of a physical state with an orthogonal physical state will give rise to at least the same experiences as the original physical state, since one can show that any computations implemented in the original state will be implemented in the superposed state. Byrne and Hall point out an apparently absurd consequence: that the same organization and experience will be present in any system not orthogonal to the original system.

I have to eat some crow on this one: there is a mistake in the reasoning in the book. The mistake (first pointed out to me by [Jacques Mallah](#)) is a technical one in the discussion on p. 350 arguing

that a superposition implements the original computation. I establish a mapping between states of the superposed system and the original implementing states of the original system by projection onto the hyperplane of the original system. I then say that since the Schrodinger equation is linear, the state-transitions among the superposed states as defined will mirror the transitions among the original states. But this last step does not follow. In effect, it presupposes that projection commutes with Schrodinger evolution, which is false in general.

How to fix this up? One needs to find a special class of superpositions, with associated projections (or other relevant operators) onto subspaces, such that within this class, projection commutes (or approximately commutes) with Schrodinger evolution. This is what's needed to capture the Everett-style idea that evolution in distinct "components" of the superposition will sometimes be more or less independent. I think that ideas drawn from decoherence theory may be useful here -- so the result won't hold for superpositions in general, but will hold for "decoherent superpositions". This makes for some interesting links with other ideas, but I haven't yet worked out all the details. If you have any ideas about this, please let me know!

Byrne and Hall have some other objections to my treatment, but I think that this is easily the most serious. I think they mislocate the source of the absurd consequence in a principle about causal relevance (ultimately stemming from preferred bases), rather than in the mistake above, which they don't discuss. If the above problem can be fixed (e.g. via the decoherence idea), I think all the causal relevance one needs will follow, on a natural treatment of causal relevance. And the independence of evolution in the subspaces will deliver all the "preferredness" of bases that one needs.

Paul Churchland. The rediscovery of light. *Journal of Philosophy* 93:211-28, 1996.

This parodies various antireductionist arguments, including my "hard problem"/"easy problem" arguments, by giving analogous arguments that someone might have given about "light" a couple of centuries ago. A response is [here](#) (contained within my JCS response).

William Greenberg. On Chalmers' principle of organizational invariance and his 'dancing qualia' and 'fading qualia' thought experiments. *Journal of Consciousness Studies* 5:53-58, 1998.

Greenberg makes the same sort of point against my "dancing qualia" argument as that made by van Heuveln et al (below): i.e., because there are two "instantiations" of the system, there is no "inner eye" before which the experiences dance. My response is similar: I think the existence of a single subject is enough to make the argument go through. Greenberg responds to my discussion of this in the book by granting that the systems might have an experienced *sense* of personal identity, and will *report* continued personal identity, but noting that that's not enough. I agree that's not enough, but this misses the point of making the change small. Presumably the "sense" and the report will also be preserved in a 75% replacement, but actual personal identity won't be; whereas replacing a small subsystem seems enough to preserve not just sense and report but

identity.

Greenberg also objects to my talk of "tepid pink" as the consequence of "fading" red qualia, but obviously this is quite inessential to the argument. And he says the argument assumes that functionally equivalent silicon replacement of neurons is possible, but that isn't so; see my [JCS response \(#3.5\)](#), for more on that.

***Christopher Hill. [Chalmers on the a priority of modal knowledge](#). *Analysis* 58:20-26, 1998.**

Hill notes that the arguments in my book depend on the claim that we have *a priori* access to logical possibility of statements according to their primary intensions, and suggests that my arguments for this claim in Chapter 2 beg the question, because they assume *a priori* access to the space of logically possible worlds (which Hill disputes).

I think Hill misunderstands the passages he discusses. At this point (p. 68) I have argued that the two-dimensional framework handles all the standard Kripkean *a posteriori* necessities, with just one space of worlds (the conceivable worlds) and two intensions. The standard examples thus give no reason to deny that conceivability of a world (as opposed to conceivable truth of a statement) implies possibility. I then *stipulate* that a "logically possible" world is an (ideally) conceivable world, noting that any reasons for believing in a narrower class of "metaphysically possible" worlds must be independent of the standard reasons. Given this stipulation, the claim about access to logical possibility that Hill finds question-begging is relatively trivial.

The real locus of disagreement is whether the class of metaphysically possible worlds is narrower than the class of logically possible worlds (so construed). This issue is discussed in the section on "strong metaphysical necessity" in Chapter 4; the passages Hill discusses in Chapter 2 are irrelevant. For more clarification of the structure of my argument here, see "[Materialism and the Metaphysics of Modality](#)", section 3.1.

Christopher Hill. Imaginability, conceivability, possibility, and the mind-body problem. *Philosophical Studies* 87:61-85.

This interesting paper argues that there is a cognitive explanation for our conceivability intuitions about consciousness (in terms of different modes of access to physical and phenomenal states), so that these conceivability intuitions don't imply anything about metaphysical possibility or about ontology. I address this matter in [section 3.4 of "Materialism and the Metaphysics of Modality"](#), in which I respond to a related paper by Hill and McLaughlin.

***Jenann Ismael. Science and the phenomenal. *Philosophy of Science* 66:351-69, 1999.**

Ismael argues that phenomenal properties are ostensively identified properties of physical space,

used to "interpret the map" associated with physical theories. The epistemic gap between physical and phenomenal in effect stems from the gap between knowing the objective "co-ordinates" of the property and knowing that the property is the ostended one (analogous to the gap between knowing the co-ordinates of a location and know that that location is *here*).

The philosophy of science here is very interesting, but all the argumentative work against me is done by the analogy with indexicals such as "I" and "here". I discuss this analogy in the book, but it's a popular strategy (e.g. it plays a role in [John Perry's](#) recent work), so it may be useful to say more here. I think the analogy can't work, for a number of reasons. First, indexical information "disappears" from the third-person point of view: an objectively omniscient third party has no significant epistemic ignorance regarding any indexical facts about another. But phenomenal information does not: a physically omniscient third party may be epistemically ignorant of phenomenal facts about another. So phenomenal information is "objective" in a way that indexicals are not. This can also be brought out by the fact that the identity between indexically identified states and phenomenal states is cognitively significant, as discussed in ["The Content and Epistemology of Phenomenal Belief"](#).

Indexical ignorance always stems from ignorance of "locating information" concerning the identification of certain objective parts of the map with "I" and "here" and "this". (We can think of this information as coming in the forms of "arrows" pointing to certain qualitative parts of the map, giving information about the "center" of a centered world.) Once an agent has all the qualitative information and all the locating information, they're in a position to know everything. But we can give Mary, or any third party, all the locating information we like about their own location in the world -- even information about the phenomenal states they ostend, if you like (though there are tricky issues here) -- and they'll still be ignorant about phenomenal states of others (what it's like to see red, what it's like to be a bat). So locating information cannot cross the relevant epistemic gap.

Ismael handles the zombie scenario by noting that imagining C-fiber firings (etc) that no-one is conscious of is just like imagining bits of space that are not "here" for anybody. But I think this point backfires: the space scenario (unlike the zombie scenario) involves a change in objective physical matters. Given all the physical information, one will know whether or not the bit of space is "here" for anyone (modulo worries about consciousness); but one will not analogously know whether the physical states are "conscious" for anyone. As before, there's no epistemic gap for facts about the objectified "x is at P" relation, but there is for the objectified "x is conscious of Q" relation.

***Mark Johnston.** [Manifest kinds](#). *Journal of Philosophy* 94:564-83, 1997.

This article mostly argues that the water-H₂O relation (and such) should be seen as constitution rather than identity. An appendix applies the discussion to issues about the explanatory gap raised by Frank Jackson and me. He argues that when things are put in terms of constitution rather than

identity, the problems fall away. For example, the identity of pain with C-fibers doesn't follow *a priori* from a story about C-fibers playing a relevant causal role, but the constitution of pain by C-fibers does.

In response: first, identity plays no special role in my arguments. All that matters is that the macro facts are entailed by the micro facts. In fact I'm happier with talk of constitution than with talk about identity here. Johnston thinks the arguments don't go through when constitution is invoked, as he claims that "if C-fibers explain the causal powers of pain, then C-fibers constitute pain" is *a priori*. Response: whether or not this is *a priori*, it is not the relevant conditional for reductive explanation. What matters is entailment of the pain facts by the physical facts. Johnston's conditional can't ground such an entailment, as it builds claims about "pain" into the antecedent. For it to do the job, we would have to replace "the causal powers of pain" by a list of specific, neutrally characterized causal powers (e.g., being caused by tissue damage, causing avoidance behavior, etc.). But then the modified conditional clearly will not be *a priori*, as it's epistemically possible that any such list be satisfied without pain. In the cases of "water" and such, by contrast, it's much more plausible that there is such a list that makes the analogous conditional *a priori*. Johnston says that derivability may fail even in other cases (slipperiness, redness, wetness), as we can't derive facts about appearance *a priori*. I agree with this, but think that this failure is a consequence of the underivability of facts about consciousness. As I say in the book, those facts are still derivable "modulo consciousness".

Robert Kirk. Why there couldn't be zombies. *Aristotelian Society, Supplementary Volume 73*:1-16, 1999.

Kirk is a onetime zombiephile turned zombiephobe. He argues that (1) on the zombiephile view, we are compounds of a physical zombie and nonphysical qualia, (2) a zombie can't detect nonphysical qualia, and qualia can't detect nonphysical qualia, (3) their compound (i.e. us) can't detect nonphysical qualia, either, so (4) the zombiephile view has the absurd consequence that we can't detect qualia.

I could take issue with (1), but I think the deepest problems are with the second conjunct of (2) and with (3). I think there's a case to be made that our phenomenology alone implies knowledge of qualia; e.g. a disembodied being with exactly my phenomenology would know it had qualia. If so, the second conjunct of (2) is in effect false. The problem is that Kirk assumes a "thin" notion of the phenomenal on which it implies nothing cognitive; I am increasingly inclined to think this is wrong. Further, the step from (2) to (3) is invalid: it's simply false that if neither A nor B can know that P, then a compound of A and B cannot know that P (e.g., on a physicalist view, split the brain in two jagged halves). So even if a zombiephile accepts (2), they might reject (3), presumably by holding that phenomenal knowledge is supported by some combination of a physical and a phenomenal state (e.g. by a physical judgment in combination with acquaintance with a quale, as I suggest in the book). So the argument can't do the work it needs to do.

Noa Latham. Chalmers on the addition of consciousness to the physical world. *Philosophical Studies*, forthcoming. ([Abstract](#))

Latham gives a number of arguments against my view that consciousness supervenes naturally but not logically on physics. His first main argument is that the view requires an extraordinarily large number of new psychophysical laws, at least one for every basic sensory quality, and possibly more. I think this is thinking about the laws the wrong way. Rather than a swarm of separate laws, one could (for example) imagine a relatively simple transformation operator from physical space to phenomenal space. These spaces may be high-dimensional, and the collection of the projections onto individual dimensions (which seems to be what Latham is looking at) may be quite complex, but this is consistent with the overall mapping being very simple; witness simple mappings between infinite-dimensional spaces in mathematics. It may also be that the dimensionality of phenomenal space can be reduced considerably by constructing it from a few protophenomenal properties. (Here Latham responds that we don't know what these properties are; but of course this is not to say they don't exist.)

Latham also considers cases such as pleasure/pain inversions (keeping physical facts constant), and related cases where experience become "inappropriately" matched with functional role. Latham argues that if these case were logically possible (as they are on my view), then we could not rule out the epistemic possibility that they are actual and that we are ourselves inverted in this way. After all, what could our certainty that we are not so inverted be grounded in? It could not be grounded in our acquaintance with the experiences, as this acquaintance is also present in the inversion cases, and it couldn't be in our beliefs, for the same reason.

In response: it's not obvious to me that the notion of "appropriateness" is entirely clear and absolute (as opposed to a product of familiar association), and it's also not obvious that we are *certain* we are in an "appropriate matching" state. (I'm certain when I am in pleasure, but I am not certain about functional roles in my brain and behavior.) But even conceding those points, I don't see that the logical possibility of inversion rules out my knowledge that I am not inverted. It is plausible, for example, that the justification of our beliefs about experiences lies not just in the experiences, and not just in the beliefs, but in the relations between the beliefs and the experiences. Those relations are not present in the inversion cases, and so the knowledge may be present in our cases even though it is not present in those. It's an interesting issue to investigate further, though.

Latham also adapts my own "dancing qualia" arguments to suggest that these are not just naturally but logically impossible. He thinks we can be certain our experience hasn't changed radically in the very recent past (whereas we couldn't be certain if dancing qualia were logically possible); I think we can't be certain of this any more than we can be certain that the world wasn't created in the very recent past. Latham suggests that although he can't rule out zombie worlds by this sort of strategy, the strong logical constraints (of functioning on phenomenology) he has established give indirect evidence for logical determination. I think this is a nice strategy, but that all the

arguments for logical constraints go wrong.

***Joseph Levine. Review of *The Conscious Mind*. *Mind* 107:877-881, 1998.**

Although a review, I mention this here because it is substantial and because it makes a point related to Melnyk's point below. Levine also challenges my arguments on the grounds that primary intensions need not be *a priori* accessible. He has two arguments for this: (1) infants and animals may have concepts without having any sort of access to their primary intensions; (2) certain contemporary theories of content (e.g. nomic covariation theories) suggest that concept's application-conditions may be not be cognitively accessible in this way, because they have "non-ascriptive" modes of presentation.

In response: (1) the infant/animal case isn't a counterexample to the *a priori* accessibility claim, any more than the fact that animals can't do mathematics shows that mathematics isn't *a priori*. Cases turning on cognitive limitations are never counterexamples to a claim of *a priori* knowability. (2) I think this is a better argument against those theories of content than it is against the accessibility claim. Levine suggests that the application-conditions of "water" and "chair" may be inaccessible in this way, but he does not address the obvious point (made in the book) that given qualitative knowledge of the external world, we are in a position to know what "water" and "chair" refer to. I take it that this is a datum that any theory must either take into account or else provide substantial arguments against; I have seen no such arguments in the literature on these theories. So it seems to me that the accessibility claim is not threatened. At best, Levine needs to take the "consciousness" concept as a special case.

I also think that even if Levine is right about inaccessibility, this is not enough to threaten the anti-materialist argument. This sort of inaccessibility would not deliver "strong necessities" (which are what the materialist needs), but at best "inscrutable truths". See ["Does Conceivability Entail Possibility"](#) for the distinction.

Harry Lewis, Consciousness: Inexplicable - and useless too? *Journal of Consciousness Studies* 5:59-66, 1998.

Lewis makes a few criticisms of the view set out in my ["hard problem" paper](#), arguing that the problem is still harder than I say, mostly because I don't take adequate account of the ineffability of consciousness. He doesn't like my listing "awareness" and "reportability" among the easy problems, for reasons bound up with this; I clarify my usage here in my [JCS response \(#3.1\)](#). There are a few places where he seems to misplace my view, but I've clarified those things elsewhere. Lewis's central point is that the ineffability of experience precludes a theory bridging processing to experience, but this is far from clear to me, and he gives little explicit argument for this conclusion. It may be that we can develop some sort of phenomenological formalism (geometric, informational, mathematical, linguistic?) within which experiences can be described and taxonomized. We already know we can do this for some aspects of experience (the more

obvious structural aspects), and I think it is an open question how far we can take it. Certainly this is an area that needs to be investigated carefully in coming years.

***Andrew Melnyk. Chalmers' argument for dualism about phenomenal properties. (APA presentation.)**

Melnyk argues against the move from conceivability to possibility. In particular he argues that the conceivability of P does not entail the existence of a possible world on which P's primary intension is true. He does this by arguing that the primary intension of a concept need not be known *a priori* by a user of that concept. He points to concepts such as "water" to suggest that I can have these concepts without knowing their (primary) application-conditions.

I agree that one can have a concept without "knowing" its primary intension in the sense of being able to articulate explicit conditions characterizing it. But this is a much stronger claim than I need. All I need is the claim that a primary intension is *accessible*, i.e. that I have *a priori* access to the relevant function from worlds to referents, i.e. that given a specification of a world, I am in a position to determine the referent. This is a much weaker claim, both because it doesn't require explicit definition, and more importantly because *a priori* knowability is much weaker than automatic knowledge (a lot of *a priori* truths are highly nontrivial). It seems to be a fact about us that given knowledge of the actual world, we're in a position to assess reference and truth-value (see pp. 57-60 of the book); this suggests that the *conditional* from world to reference/truth-value is knowable, prior to knowledge of how the external world actually turns out. And nothing in Melnyk's discussion suggests otherwise. At least, it is the *a priori* accessibility claim that Melnyk needs to challenge. (I believe Melnyk is revising the paper to take this sort of point into account.)

John Perry. *Knowledge, Possibility, and Consciousness*. MIT Press, 2001.

Perry discusses my work at a number of places in this book. I have written a response (destined for a symposium in *Philosophy and Phenomenological Research*): ["Insentience, Indexicality, and Intensions"](#).

Robert Stalnaker. What is it like to be a zombie? In (T. Gendler & J. Hawthorne, eds) *Imagination, Conceivability, and Possibility*. Oxford University Press, forthcoming.

I reply to the main argument in this paper at the end of ["Does Conceivability Entail Possibility?"](#).

***Robert Stalnaker. [On considering a possible world as actual](#). *Proceedings of the Aristotelian Society*, Supplementary Volume, 2001.**

This discusses two-dimensional semantics as a way to analyze Kripkean a posteriori necessity. Stalnaker defends his own "metasemantic" account and argues against the sort of "semantic"

account given by me and others. I think the metasemantic account is pretty clearly unsatisfactory, as it can't capture the distinction between a priori and a posteriori necessities. His objections to the semantic account are brief and inconclusive, but he appears to commit his opponent to a version on which primary intensions are evaluated by finding the extension of a token with the same meaning as the original token, placed in the new context. I think that's the wrong way to do things: see the discussion of "epistemic" and "contextual" interpretations of 2-D semantics in ["The Foundations of Two-Dimensional Semantics"](#).

Nigel Thomas. [Zombie killer](#). In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press, 1998.

Thomas argues that the concept of a zombie is incoherent, by appealing to the fact that zombies will claim to be "conscious". He considers the possibility that we might interpret this claim as true, false, or meaningless, and argues that trouble threatens each way.

I discuss this sort of thing at length in Chapter 5 of my book, where I argue that the zombie's claim is best taken to be false (though I have occasional sympathy with the "meaningless" option), e.g. because the zombie's "consciousness" concept is empty, a bit like "phlogiston". So the zombie is mistaken. In response, Thomas notes that the zombie's claim is formed by the same "cognitive process" as an ordinary conscious being's claim, so if the zombie is mistaken, this cognitive process is unreliable, throwing strong doubt on our own claims that we are conscious. This line of reasoning was anticipated in the book, where I suggest that the justification for my belief that I am conscious lies not just in my cognitive mechanisms but also in my direct evidence; the zombie lacks that evidence, so his mistake does not threaten the grounds for our beliefs. (One can also note that the zombie doesn't have the same beliefs as us, because of the role that experience plays in constituting the contents of those beliefs.) Thomas's discussion, although it raises some intriguing issues, never addresses this natural line of defense.

Bram van Heuveln, Eric Dietrich, & Michiharu Oshima. Let's dance! The equivocation in Chalmers' dancing qualia argument. *Minds and Machines* 8:237-49, 1998.

This paper tries to escape the dancing qualia argument by on the grounds that the argument assumes that one system is present throughout, but it may be the case that on switching subsystems two different systems come into and out of existence. So really two distinct 'phenomenal worlds' are present, and it's not surprising that no change is noticed.

What really matters here is whether two distinct *individuals* (i.e. subjects) are involved. If two distinct individuals are involved, the argument won't go through. I tried to handle this point in the book by making the changed area small enough (less than 10 percent of the brain) that it was plausible that the same individual was present throughout. It seems unlikely to that replacing some small part of the visual cortex (say) will be enough to change my identity. If one thinks 10 percent is too much, we can even make the replacement 5 percent or less, without too much cost.

The authors don't really address this point. I think they hold that the mere presence of two physically distinct subsystems is enough to make their point, but I don't see this. If we grant that all this involves changes in *my* consciousness, as seems plausible, I think the argument goes through.

The purported "equivocation" is one between 'change in experience' and 'experience a change'. They argue that although the experiences changes, no single individual experiences a change. I do require that a single individual experiences a change, at least in the sense that the experiences that change are those of an individual, although not in the sense that the individual notices the change, which obviously can't happen in this set-up. The point of the small replacement is precisely to make it plausible that this happens.

***Stephen Yablo, [Coulda, woulda, shoulda](#). In (T. Gendler & J. Hawthorne, eds) *Imagination, Conceivability, and Possibility*. Oxford University Press, forthcoming.**

This paper is a critique of two-dimensional semantics and my use of it in establishing a link between apriority and possibility. It's a rich paper with a number of different arguments. I reply to most of them in "[Does Conceivability Entail Possibility?](#)", which is forthcoming in the same volume. Note that my responses are interspersed with the main text, at three different locations in the article: [here](#), [here](#), and [here](#). The second of these also serves as a response to Yablo's "[Modal Rationalism and Logical Empiricism: Some Similarities](#)", and the third serves as a response to his "[Grokking Pain](#)". See also our earlier exchange in *Philosophy and Phenomenological Research*, referenced [above](#).

Go to:

- [David Chalmers' home page](#)
- [Online discussions of my work](#)
- [Reviews of *The Conscious Mind*](#)

THE TYRANNY OF THE SUBJUNCTIVE

David Chalmers
Idaho, March 24, 2000

(1) A CONTRAST BETWEEN INDICATIVE AND SUBJUNCTIVE CONDITIONALS

- (1a) If Prince Albert Victor killed those people, he is Jack the Ripper (and Jack the Ripper killed those people).
- (1b) If Prince Albert Victor had killed those people, Jack the Ripper wouldn't have (and Prince Albert wouldn't have been Jack the Ripper).
- (2a) If XYZ is the liquid in the oceans and lakes, water is XYZ (etc).
- (2b) If XYZ were the liquid in the oceans and lakes, water wouldn't be in the oceans and lakes (etc).

The above all seem to be intuitively correct, even though (1a) and (2a) have "metaphysically impossible" consequents.

What's the relevant difference? Indicatives take the antecedent as a hypothesis about the actual world, whereas subjunctives take it as a description of a way the world might have been but actually isn't (Fowler: "it belongs to utopia"). That is, indicatives consider the antecedent as actual; subjunctives consider it as counterfactual.

In terms of the 2-D semantic framework: evaluation of indicatives tracks the primary intensions of the expressions involved, whereas evaluation of subjunctives tracks their secondary intensions. "Metaphysical necessity" reflects secondary intensions, not primary intensions.

(2) THE ANALYSIS OF NECESSITY AND POSSIBILITY

S is necessary $\leftrightarrow$ S is true in all possible worlds.
S is possible $\leftrightarrow$ S is true in some possible world.

QUESTION: How to evaluate "S is true in W"?

SUBJUNCTIVE READING: "S is true in W" iff

If W were the case, S would be the case.

INDICATIVE READING: "S is true in W" iff

If W is the case, S is the case.

The first tracks familiar "subjunctive" (utopian) evaluation of possibilities. The second tracks "indicative" (epistemic) evaluation of possibilities. (Is the epistemic possibility that W is actual an instance of the epistemic possibility that P?)

There are two resulting sorts of "necessity": subjunctive necessity and indicative (epistemic) necessity. And two sorts of "possibility", etc.

Plausibly: S is subjunctively necessary if S's secondary intension is true in all possible worlds. S is indicatively necessary if S's primary intension is true in all (centered) possible worlds.

Primary intension: is S true of W considered as actual?
(evaluated in the way in which we evaluate epistemic possibilities)

Secondary intension: is S true of W considered as counterfactual?
(evaluated as we evaluate explicitly counterfactual possibilities)

e.g. S = "water is H₂O". W = the Twin Earth world (with XYZ).

(3a) If W is the case, water is not H₂O (it's XYZ).

(3b) If W were the case, water would still be H₂O (and XYZ merely watery).

This mirrors the fact that the secondary intension of S is true in W, but the primary intension of S is false in W. So "water is H₂O" is not indicatively (epistemically) necessary, although it may be subjunctive necessary.

(3) THE SYMMETRY THESIS

In contemporary philosophy (post-Kripke), "necessity" (simpliciter) is always read as subjunctive necessity, and "P is true in W" is always evaluated by the subjunctive reading.

In fact: subjunctive conditionals *ground* the contemporary evaluation of statements in worlds, and the consequent analysis of necessity. Kripke's arguments in Naming and Necessity are always ultimately grounded in intuitions about subjunctives.

QUESTIONS: Why is this? Arbitrary? Argued or stipulated? Is indicative necessity an equally good candidate for "necessity"?

SYMMETRY THESIS: Indicative necessity is as good a candidate for "necessity" as subjunctive necessity. And epistemic evaluation is just as reasonable a reading of "P is true in W" as subjunctive evaluation.

If the symmetry thesis holds, it should be just as reasonable to say that:

"Hesperus is visible in the evening (if it exists)" is necessary;
"water is H₂O" is contingent;
"I am here now (if I exist)" is necessary;
"the meter stick in Paris is one meter long (if it exists)" is necessary;
the necessity of identity doesn't hold;

and

In the XYZ-world, water is XYZ;
It's not necessary that I am David Chalmers;
In some possible worlds, Cicero is not Tully;
etc.

(4) SIX POSSIBLE REASONS FOR FAVORING THE SUBJUNCTIVE

(a) Indicative necessity is "merely epistemic".

[Answer: So? Before 1970, almost everyone thought necessity was tied to the epistemic (cf. Pap's book). Kripke *argued* that necessity and epistemic notions came apart, by appeal to the subjunctive, but

one can't simply presuppose it.]

(b) S is necessary $\leftrightarrow$ S "could not have been otherwise" -- subjunctive!

[Answer: (i) "Could not have been otherwise" is ambiguous between epistemic/subjunctive readings. (ii) This reading doesn't seem to have been in wide currency pre-1970.]

(c) Indicative necessity requires centered worlds.

[Answer: Possible worlds are a technical device, and there doesn't seem to be anything in the notion of necessity itself to bar an indexical element.]

(d) Subjunctive necessity supports quantified modal logic, whereas indicative necessity doesn't.

[Answer: True enough. But whether necessity allows quantifying in was a live issue -- it can't be a prior constraint on necessity that it does! (Though one can see why Kripke would be attracted to a notion of necessity that does.) At best a split verdict: Quine right about indicative necessity, Kripke about subjunctive necessity. cf. Burgess.]

(e) The relevant indicative conditionals aren't true/false, just assertible, due to epistemic relativity. (e.g. Sly Pete)

[Answer: This epistemic relativity doesn't affect epistemic evaluation of statements in worlds. At most it affects the move from incomplete antecedents to relevant worlds.]

N.B. Possible-worlds analysis of indicative conditionals:

"If P, then Q" is correct/true/assertible $\leftrightarrow$

The epistemically closest P-world(s) is a Q-world

(where here a P-world is a world satisfying the primary intension of P).]

(f) Some indicative "necessities" will vary in truth-value between users, as primary intensions of some terms (e.g. names) can vary between users.

[Answer: True, but the same is the case for subjunctive necessities, e.g. "I am DC", as secondary intensions of some terms (e.g. indexicals) vary between users. At worst, bring in a sentence-meaning/utterance-content distinction.]

(5) THE TYRANNY OF THE SUBJUNCTIVE

Necessity plays a central role in almost every area of philosophy. So the analysis of necessity affects almost every area of philosophy. Kripke's analysis has been central in supporting essentialism in metaphysics, direct reference views in the philosophy of language, externalism in the philosophy of mind, and so on. Has led to a lot of interesting remaking of philosophical views, with "marvelous internal coherence". But.

Imagine an alternative universe in which Kripke* used indicative conditionals instead. His book *Naming and Contingency* might have

* defended a strong link between the a priori and the necessary (plausibly S is a priori iff it is indicatively necessary)

* have argued against the necessity of identity and de re necessities (identities aren't indicatively necessary, and indicatives don't allow substitution or quantifying in, so de re necessities don't make sense [except for numbers?]).

- * have argued for a Fregean view of meaning and content (since primary intensions can plausibly play the role of sense -- they both determine reference and reflect cognitive significance).

Imagine that the ensuing philosophical analysis of necessity had been as biased toward indicative necessity as recent philosophy has been toward subjunctive necessity. The resulting dominant views in various areas of philosophy might have been very different:

- * philosophy of language: Fregean views, not direct reference.
- * philosophy of mind: internalism, not externalism
- * metaphysics: nominalism, not essentialism; etc.

Neither our world nor W^* is the right outcome. The best outcome, given the symmetry thesis, is one in which the notion of "necessity" is recognized as ambiguous. Kripke** (importantly) recognizes the modality of subjunctive necessity and its distinctive properties. This leads the pretheoretical notion of "necessity" to split into two varieties, along with two corresponding means of evaluation in possible worlds. As to specific philosophical areas: the outcomes here will depend on analyzing whether the modality in those areas is relevant qua subjunctive or qua epistemic.

- * Metaphysics: A split verdict about essentialism, with Quine right about indicative necessity but essentialists right about subjunctive necessity. But "subjunctive essentialism" is a significant sort of essentialism (it's significant that there are ways an entity could not have been) -- so the door is reopened to essentialism in philosophy.
- * Philosophy of mind, epistemology: Here it's plausible that the central role of modality by far (at least where belief content is concerned) is epistemic -- e.g. considerations of rational inference, cognitive significance, and so on. And that the most relevant application conditions of concepts to worlds are epistemic, not subjunctive. So the intensional analysis of mental content ends up internalist. (There are still issues about externalist belief attribution to handle, but they can be handled straightforwardly.) As a bonus, we have a notion of belief content that is much more useful in epistemology.
- * Philosophy of language: A split verdict? Expressions have complex semantic value, with something like a primary intension (more or less Fregean) and a secondary intension (more or less referential, for some terms). The latter are needed to analyze subjunctive conditionals and maybe belief attributions (though: scope analysis?). The former are needed to analyze indicative conditionals, belief attributions, etc; and to analyze the rational connections among statements, and to analyze the connection between language and thought. The modality inherent in Frege's notion of sense is arguably epistemic (sense determines the epistemic application-conditions of a concept), so Fregean sense is vindicated as above (modulo the introduction of an indexical element and the occasional variation of sense between tokens).

Is there synonymy in Ockham's mental language?

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

[[Published in *The Cambridge Companion to Ockham*, edited by Paul Vincent Spade. Cambridge University Press, 1999.]]

1 Introduction

William of Ockham's semantic theory was founded on the idea that thought takes place in a language not unlike the languages in which spoken and written communication occur. This mental language was held to have a number of features in common with everyday languages. For example, mental language has simple terms, not unlike words, out of which complex expressions can be constructed. As with words, each of these terms has some meaning, or signification; in fact Ockham held that the signification of everyday words derives precisely from the signification of mental terms. Furthermore, the meaning of a mental expression depends directly on the meaning of its constituent terms, as is the case with expressions in more familiar languages.

As one might expect, there are important differences between mental language and everyday languages. For example, mental languages signify their objects naturally rather than conventionally. At a more concrete level, Ockham suggested that numerous features of spoken or written language - participles and pronouns, for example - might not exist in mental language.

Two ubiquitous features of everyday languages are the phenomena of equivocation and synonymy. The first of these is exemplified in English words such as 'bank', which have two entirely different meanings. The second is exemplified by pairs of terms such as 'bachelor' and 'unmarried man', which share a common meaning. The question arises: are these features also found in mental language? It seems to be commonly accepted that they are not. Ockham himself is not entirely clear on the matter, but Trentman and Spade have argued, based on both textual and theoretical considerations, that the most coherent position broadly compatible with Ockham's work is that neither synonymy or equivocation may occur in mental language.[1] I shall not discuss the status of equivocation in this paper, as I find the arguments on that topic persuasive. However, I wish to argue that the case against synonymy is not as strong.

My argument for synonymy is largely theoretical, although it also has a textual element. The theoretical

case has a positive and a negative part. On the positive side, I will argue that there are a number of reasons why mental synonymy might exist, and that a mental language without synonymy would be relatively clumsy, with a number of ad hoc features. On the negative side, I will address various arguments that have been put forward against the possibility of synonymy, and will try to show that they are not conclusive. Textually, I will argue that although Ockham appears to deny the possibility of synonymy in mental language, he also makes remarks that commit him to that possibility. I do not think that my arguments are entirely conclusive, but I hope to demonstrate that the possibility of mental synonymy is not as implausible as has sometimes been thought.

2 Positive Arguments

2.1 Logical Equivalence

The first argument is based on the observation that once certain logical primitives are admitted into a language, then it seems to follow immediately that certain complex expressions are synonymous. Consider, for example, the English expressions 'man and (cat or dog)' and '(man and cat) or (man and dog)'. The parentheses are introduced for clarity, but they are not strictly necessary, being straightforwardly replaceable by longer locutions (for instance, 'a man with either a cat or a dog' and 'either a man with a cat or a man with a dog'). Call the two expressions E1 and E2. It seems apparent that E1 and E2 are synonymous. It is a straightforward logical inference to go from a sentence involving E1 to the corresponding sentence involving E2, and vice versa. It is plain that the two expressions come to exactly the same thing in the matter of signification.

As it is with written language, so it should be with mental language, at least in this case. Mental language presumably has among its simple terms such logical operators as 'and' and 'or'; if it does not have these, then it presumably possesses others with equivalent power, on the basis of which a similar case could be made. There is no doubt that simple expressions such as 'dog' and 'cat' exist in mental language (if not these, then any simple nouns will do). And there seems to be no reason why these expressions and logical operators should not be combined to form complex expressions. So mental language contains the expressions 'man and (dog or cat)' and '(man and dog) or (man and cat)', or other expressions that make the same point. These expressions, it seems, must be synonymous, for just the reasons given above. Therefore there is synonymy in mental language.

There are two paths that an opponent of mental synonymy might take in countering this argument. She could argue that the two expressions are not in fact synonymous, or she could argue that the two expressions are in fact identical in mental language, so that no synonymy between different expressions need be introduced. I will consider these objections in turn.

2.1.1 The First Strategy

In the *Summa logicae*, [2] Ockham says that for two terms to be synonymous, they must signify exactly

the same things, and they must signify these *in the same way*. An opponent of mental synonymy might argue that although the two expressions given above signify the same things, they signify them in different ways, and so are not synonymous.

It is not entirely clear what Ockham meant by 'to signify in the same way'. Spade argues that two mental expressions signify in the same way if and only if they are syntactically equivalent - that is, if and only if they consist of exactly the same categorematic expressions in exactly the same syntactic constructions.[3] If this were the case, then our two given expressions would certainly not be synonymous. However, it seems to me that this construal of synonymy buys the conclusion entirely too cheaply, by defining mental synonymy out of existence, and I will argue against it below. For now, let us consider a weaker construal of 'in the same way'. This construal captures our intuition that what is required of synonymy is not just that two terms signify the same things, but that they *must* signify the same things. To be more precise, under this construal of synonymy, two terms are synonymous if their equivalence is *a priori* and necessary.

This yields a strong criterion for synonymy. The coextensive pair of terms "renate" and "cordate" fail the test, for example, as they are not necessarily coextensive. The coextensive pair of terms "Hesperus" and "Phosphorus" fail the test, as they are not *a priori* coextensive. It is difficult to know whether this criterion corresponds to Ockham's intentions, but it seems to capture a large part of our intuitions behind what is meant by 'synonymy', and in particular to capture the extra strength carried by synonymy compared to mere coextensiveness. And the logically equivalent pair of terms given above clearly satisfy this criterion.

(One might also invoke a criterion according to which synonymous terms are those that are substitutable *salva veritate* even in modal contexts. I do not use this criterion here, as it yields counterintuitive results when combined with the contemporary understanding of necessity: for example, "Hesperus" and "Phosphorus" come out synonymous. But on the medieval understanding of necessity, this criterion may yield the same results as the criterion given above, and less anachronistically.)

An opponent might try to argue that this criterion is not strong enough: although the logically equivalent pair of terms satisfies the criterion, the terms nevertheless signify in subtly different ways. For example, our opponent might argue that the two terms are not intersubstitutable in certain epistemic contexts, and are therefore not synonymous. For example, it might be the case that John believes that to gain admission to the party, one must be accompanied by a man with either a dog or a cat; but John does not believe that to gain admission to the party, one must be accompanied by either a man with a dog or a man with a cat. The reason for this, presumably, would be that John is not very capable at logic, and so has not made the straightforward inference. Now, we might respond to this by arguing that in fact, if John has the first belief, he has the second belief whether he knows it or not. He believes it implicitly, one might say. But for the sake of argument let us accept that John has the first belief but not the second. If this is so, then it is apparent that in this epistemic context, the two terms are not intersubstitutable.

It seems to me, however, that intersubstitutability in such epistemic contexts is too strong a condition to

require for synonymy.[5] If we admit this as a criterion, it would seem to follow that there are no synonyms at all. For given any pair of purported synonyms - e.g., 'bachelor' and 'unmarried man' - we can construct examples like those given above. For example, John might believe that all bachelors are invited for dinner, but might not believe that all unmarried men are invited for dinner - presumably because he is once again too muddled to make the required inference. (Again one might argue that if John has the first belief, he has the second whether he knows it or not. But as we disallowed this line of reasoning for the sake of argument in the paragraph above, we must equally disallow it here.) This conclusion shows that the suggested criterion is too strong. In fact, it seems that even Ockham himself would reject intersubstitutability in epistemic contexts as a criterion for synonymy, as the following passage demonstrates[6]:

Those synonyms are broadly so called even though not all users believe them to signify the same [thing] but rather, under a deception, they judge something to be signified by the one that is not signified by the other.

It follows that this argument provides no reason to reject the synonymy of E1 and E2, or at least no more reason than there is for rejecting the synonymy of such terms as 'bachelor' and 'unmarried man'. Logical synonymy seems to be at least as strong as common-or-garden conceptual synonymy, so that if we want to retain the notion of synonymy at all, we had best admit the possibility of logical synonymy, at least in written language.

Once we have accepted the possibility that the two written expressions E1 and E2 are synonymous, then it follows that either (a) they are subordinated to synonymous mental expressions, thus establishing my case, or (b) they are subordinated to identical mental expressions, thus saving the same for the opponent of mental synonymy. I will consider possibility (b) next.

2.1.2 The Second Strategy

An opponent's second strategy is to argue that only one expression in mental language is involved here: the two mental expressions that we are labeling 'man and (cat or dog)' and '(man and cat) or (man and dog)' are in fact the same expression. It might be claimed that the fact that they look like different expressions is an artifact of English, not mental language: the different *written* expressions E1 and E2 are in fact subordinated to the same *mental* expression. To get an idea of how this argument might run, consider a simpler pair of expressions: 'man and dog' and 'dog and man'. For reasons similar to those given above, these expressions seem to be synonymous. But an opponent of synonymy could very plausibly argue that these are not different expressions in mental language at all, as the only difference between them is one of word order - a feature that need not be preserved in mental language.[7] For example, it could well be argued that the term 'and' in mental language is not interpolated sequentially between two terms in a particular order, but rather is bound to them both in some symmetrical fashion, such as a tree structure, or perhaps as a common ending (something like 'man-and dog-and', where we take it that the two words are unordered). On this plausible account, the difference in the two expressions 'man and dog' and 'dog and man' would not persist into mental language.

It is difficult to see, however, how we could apply an argument like this to more complex cases, such as our 'man and (dog or cat)' vs. '(man and dog) or (man and cat)'. These two expressions are not just superficially different; they are structurally different. One of them is a conjunction; the other is a disjunction. However we represent 'and' and 'or' in mental language, they will need to be put together hierarchically on occasion, and it seems clear that the differences in their hierarchical order must be represented in mental language.

One way out for our opponent might be to argue that logical expressions in mental language must be reduced to some common form such as disjunctive normal form. It would then be the case that any two logically equivalent terms or propositions would be represented identically. The trouble with this is that is an ad hoc restriction that places limits on how mental terms may be combined. On the face of it, would take it that if 'man', 'dog', 'cat', 'and' and 'or' are admissible mental terms, then 'man and (dog or cat)' would be an admissible expression in mental language; but according to the present proposal this expression would somehow be debarred from being formed, for no apparent reason other than to preserve a theoretical claim. Indeed, to place such a restriction on combination of terms would seem to reduce mental language's claim to being a language in the first place, one of the key properties of language being that complex expressions can be generated recursively and compositionally, whereas according to this proposal, while 'man' and 'dog or cat' would be valid mental expressions, 'man and (dog or cat)', their conjunction, would not be. This seems arbitrary and implausible.

It seems much more plausible for mental language to consist of a set of simple terms that can be combined and recombined without any limits other than syntactic requirements. In particular, logical operators ought to be able to combine nominal terms in any combination. But once we admit this, no matter what set of logical operators we choose, logically equivalent but non-identical expressions will be able to be formed. We might be able to get around certain superficial differences in structure, such as that in the 'man and dog' case, by proposing certain symmetries in our operators, but there will always be propositions with different structures that turn out to be logically identical. One can see this easily by noting, for example, that while there are only four logically distinct truth-functional combinations of two propositions, any compositional system will yield an infinite number of syntactically distinct complex sentences formed by combining these propositions under truth-functional operators. Logical equivalence is thus endemic in any combinatorial system.

The conclusion is that given that logical equivalence implies synonymy, it follows that synonymy will be ubiquitous in mental language, unless we put ad hoc restrictions on the manner of combination of mental terms.

Before passing to the next topic, I should respond to a natural response, based on textual considerations, that may have occurred to the reader. This response notes that according to Ockham, synonymous written expressions are subordinated to the same mental expression,[8] so that no matter what I say above, the expressions E1 and E2 *must* be subordinated to a single expression in mental language. My response to this objection is of course to point out that the arguments I am giving for the existence of synonymy in mental language are equally arguments for the rejection of the claim that synonymous written

expressions are subordinated to the same mental expression. If there is synonymy in mental language, a more plausible criterion is that synonymous written expressions must be subordinated to *synonymous* mental expressions. Of course, this means that I am going against an explicit claim of Ockham's; but I am giving a theoretical argument, after all. I will address this issue further below.

2.2 Conceptual Change

I now pass to a quite different line of argument in favor of mental synonymy. This argument stems from the observation that many terms in everyday languages, and presumably in mental language, undergo gradual changes in meaning. Many current terms do not mean exactly what they meant fifty years ago. Even within a single individual, the meaning of a term can gradually shift over time. Sometimes it may even occur that two terms that originally had quite different meanings gradually drift until they mean the same thing. Perhaps, for instance, when one was a child one used the term 'stone' only for small objects, and used the term 'rock' for large ones, but over time one's use of the term has drifted until now they are used in exactly the same way, so that the terms are synonymous. Nothing in my argument turns on the specific example - all that is required for the argument is the *possibility* of this kind of gradual drift into synonymy taking place, and this seems indisputable.

So let us say that Y and Z are two terms that start out with different meaning, but which through a continuous process over time come to mean the same thing. This might even occur without the individual within whom it is happening being conscious of the drift to synonymy. He might realize at some later point that 'stone' and 'rock' are synonymous for him, whereas they were not twenty years ago, and the drift might have occurred without his realizing it at the time. Ockham himself endorses the possibility that synonymy might go unrecognized, in the passage quoted above.

Now as these terms gradually drift toward synonymy, what is happening in the mental language? Presumably Y and Z are subordinated to mental terms Y' and Z', and Y' and Z' are also undergoing a slow drift in meaning. Now what happens on the day when Y and Z finally become synonymous? Do the terms Y' and Z' suddenly become the same term? Does one of them suddenly disappear, so that for example Y and Z are now both subordinated to Y', and Z' is gone? Neither of these possibilities are strikingly plausible. The process of change is sufficiently slow that it is hard to imagine that any sudden change could be taking place in mental language. In practice, it will be very hard to locate a precisely moment at which Y and Z become synonymous; there may be a long period of approximate synonymy before we can say with confidence that they are definitely synonymous. Are we to suggest that the mental language has hair-trigger sensitivity to the process, so that at the exact moment when the meanings come to coincide, a jump suddenly occurs?

It seems at least as plausible to argue that as the drift takes place, the two mental terms remain distinct, gradually becoming synonymous but nevertheless oblivious, at least for a while, to each other's presence. It may well be for a long period that the subject does not make the expected inferences from propositions involving Y to those involving Z, because the subject does not consciously realize that the two concepts have become synonymous for him. If the corresponding mental concepts had become identical, this

inferential gap would be harder to explain. After a time, the subject may consciously realize that the concepts are synonymous, and perhaps the mental terms will then be "fused" into one, but until then there seems no reason in principle why the terms should not be distinct.

The other possibility to be considered is that as Y and Z drift in meaning, they are subordinated to a series of different terms in mental language, Y'1, Y'2, ... and Z'1, Z'2, ... Upon every change in the meaning of Y, no matter how small, Y becomes subordinated to a new Y' in the mental language. Over time, the corresponding Y' terms gradually become more like the Z' terms, until finally they become identical. This account has the virtue of avoiding the sudden change required by the previous story, but it still seems problematic. For a start, it seems to be very extravagant with terms in mental language, postulating a large number of different terms where we previously needed only two. Further, it does not solve the problem mentioned above, of the possible inferential gap in the subject's abilities - one would think that if the corresponding concepts have become identical, any inferences from one to the other would be automatic. This seems in turn to imply that two terms cannot become synonymous without a subject recognizing that fact, which appears to contradict Ockham's own claim quoted above.

On the balance of things it seems to me that conceptual change provides a strong argument but not a knockdown argument for mental synonymy. The opposing story just given seems feasible enough that it *could* be the case, although it does raise a serious question about compatibility with Ockham's own views on the recognition of synonymy, an issue I discuss further below. However, the story that I sketched of gradual drift into synonymy without identity seems equally if not more plausible, giving more reason why mental synonymy is not an altogether unreasonable idea.

2.3 Efficiency

My final argument in favor of mental synonymy is a pragmatic one. Ockham, in his writings, give very little idea about how mental language *functions* - how it enables one to make inferences from one proposition to another, for instance, or how mental propositions help determine the actions we take toward the world. We can imagine, however, that this functioning does not come for free, but rather takes some work. Evidence for this can be seen in the fact that we find it harder to make inferences from complex propositions than from simple propositions. The more complex a proposition, it seems, the harder it is for our mental system to deal with. Therefore one might draw the conclusion that in order for our mental system to function as well as it can, mental propositions should be as simple as possible (as long as they are complex enough to express the required meaning).

Now in mental language, presumably, there are certain complex expressions that get used repeatedly.[9] In the interests of efficient functioning of the mental system, it might seem to be advantageous for it to introduce internal *abbreviations* for these complex expressions. Instead of having to deal with expressions like 'man who cuts cloth and makes suits', the system would only have to deal with the simple expression 'tailor' instead. We certainly find this kind of abbreviation useful in our external practice; I am here suggesting that it might have a role in internal functioning as well. If, as we have supposed, complex expressions are more difficult for the mind to deal with, then the systematic

replacement of these by simple expressions might allow a significant enhancement to our cognitive capacities.

Of course, there would have to be systematic links between a term and its abbreviation, so that thoughts about 'tailor' could easily lead back to conclusions about suits and cloth when necessary. But inferences involving tailors would in general be much easier to make in the above form. The only downside is that we would have introduced synonymy into mental language, but it seems to me that in this context this is only advantageous. A typical argument against mental synonymy[10] has the form "Who needs it?" - in other words, what is the point of mental synonymy if synonyms are truth-conditionally equivalent and so have the same expressive power? Any distinction between synonyms would be a difference that makes no expressive difference, and so would be unnecessary in mental language. Here, we have seen that pragmatic considerations show that there might be relevant differences that are not expressive differences. Rather, they are differences that aid mental function.

This is obviously not a knock-down argument. For a start, it may be anachronistic in focusing on the way that mental propositions are processed, a way of thinking that was not so common in Ockham's time. Nevertheless, all that is required to see the point is that complex mental propositions might be more cumbersome to deal with than simple ones, and this does not seem to be a particularly advanced consideration. Of course, the argument is only a plausibility argument demonstrating why mental synonymy might be a reasonable thing, and as such is a weaker argument than two arguments given above, but nevertheless it helps in breaking down the intuition that mental synonymy would be an entirely useless thing.

3 Textual Considerations

I now consider some textual issues. In particular, I must confront what seems to be the strongest argument against synonymy in Ockham's mental language: the fact that on at least on occasion, Ockham appears to deny that mental synonymy exists. This occurs in his Quodlibet V.812:

To the principal argument, I say that everything that is an accident of a mental term is an accident of a spoken term, but not the other way around. For some [things] are accidents of spoken terms because of the necessity of signification and expression, and they belong to mental names. Others are accidents of spoken terms for the sake of the decoration of speech (like synonyms) and for the sake of well-formedness, and they do not belong to mental terms.

Now, it is not entirely clear that this is a blanket denial of the possibility of mental synonymy. It might alternatively be interpreted as a denial that synonymy in spoken language is reflected in mental synonymy, leaving open the possibility of mental synonymy that arises in other ways. Or it might be interpreted as the denial that *certain* synonyms in spoken language - those that exist merely for the sake of decoration - are reflected in mental synonyms, leaving open the possibility that other (nondecorative) spoken synonyms correspond to mental synonyms. Furthermore, even if we accept this as a denial of the

possibility of mental synonymy, it seems to be the only occurrence of such a denial in Ockham's writing, so that even Spade (who argues against synonymy in mental language) concedes that the textual support is not as strong as it could be.¹³

In any case, my central purpose is to make a theoretical argument about what Ockham should have held, rather than about what he in fact did hold. In other words, insofar as Ockham can be seen to have denied the possibility of mental synonymy, this paper should perhaps be taken more as a criticism of Ockham than as an interpretation. Of course this is a somewhat delicate matter, as if one jettisons too much of Ockham's theory one runs the risk of not so much criticizing him as ignoring him. Nevertheless, I believe that the position I am putting forward is compatible with Ockham's overall thrust, and only requires the rejection of one or two specific claims. And I argue below that Ockham's claims on this matter appear to be inconsistent, so that any interpretation must reject some of them. So this interpretation may not be worse off than any other.

The other specific claim of Ockham that causes problems for mental synonymy is his statement on a number of occasions that synonyms in spoken or written language are subordinated to identical terms in mental language. As Spade points out,^[14] it does not necessarily follow from this that mental synonyms cannot exist; all that follows is that if they do exist, they do not have associated synonyms in spoken or written language. However, this seems too weak a ground to base a defense of mental synonymy on. For a start, *if* our mental synonyms had any written or spoken terms subordinated to them, these terms would be synonyms, in violation of Ockham's claim; and on the face of it it seems reasonable that any given mental term should at least possibly have an associated written or spoken term. If we were to retain Ockham's claim and to argue for mental synonymy, we would be committed to the existence of mental terms that could not have associated spoken or written terms, and while this is not impossible it at least seems ad hoc and unmotivated. Furthermore, my best evidence for the existence of mental synonymy^[15] came precisely from the consideration of synonymous spoken or written expressions that I argued had to be subordinated to different mental expressions.

For these reasons, it is best for the defender of mental synonymy to argue that Ockham's claim should be jettisoned along with his denial (insofar as it is a denial) of mental synonymy. The correct criterion for synonymy of spoken or written terms is that they should be subordinated to synonymous mental terms, rather than to identical mental terms. This jettisoning of Ockham's claim may reduce this account's chances of being an interpretation of Ockham, but I believe that it does not reduce its overall plausibility.

I now come to the internal tension within Ockham's text. This casts further doubt on the subordination of synonymous terms to identical mental terms, and indeed suggests that Ockham may be committed to the possibility of mental synonymy, despite what appear to be claims to the contrary. In a passage quoted earlier, Ockham says:^[16]

Those synonyms are broadly so called ... even though not all users believe them to signify the same [thing] but rather, under a deception, they judge something to be signified by the one that is not signified by the other.

Ockham goes on to point out that this is the sense of synonymy with which he is dealing for most of his text. According to this passage, then, it is possible that a user can judge that 'That object is an X' is true, and at the same time judge that 'That object is a Y' is false, even though the terms X and Y are synonymous. But this implies that X and Y cannot be subordinated to the same term of mental language! For if X and Y were subordinated to the same mental term, then all mental judgments about X's and Y's would coincide. The two propositions 'That object is an X' and 'That object is a Y' would be subordinated to identical propositions in mental language, and it is impossible that one could be judged to be true and the other simultaneously be judged to be false. Therefore this passage seems to contradict his claim that synonymous terms are subordinated to identical mental terms.

In fact, if we accept the above passage at face value, it is hard not to draw the conclusion that the terms X and Y must be subordinated to different but synonymous mental terms (as it is surely impossible that synonymous spoken or written terms are subordinated to nonsynonymous mental terms), which would directly establish that synonymy exists in mental language.

It is difficult to say how we should deal with this contradiction, or how Ockham would have dealt with it, had it been pointed out to him. He might have chosen to retract the claim made in the passage above; the opposing claim that synonymous terms are subordinated to identical mental terms is certainly made more often. On the other hand, he might have seen the possibility of unrecognized synonymy in spoken or written language as providing grounds to reject the claim about subordination of synonymous terms. In any case, the internal tension revealed here further weakens the plausibility of the claim that synonymous terms are subordinated to identical mental expressions, and indicates that the claim might be rejected without doing too much violence to the rest of Ockham's system.

Furthermore, insofar as Ockham's own claims are inconsistent, it follows that any consistent interpretation must reject one or more of these claims. So Ockham's apparent denials of mental synonymy do not provide any overwhelming reason to reject the view I have offered in favor of any other consistent interpretation. This opens the way to accepting the possibility of synonymy in Ockham's mental language.

4 Other Negative Arguments

The central negative argument against the possibility of mental synonymy is the textual argument I have just discussed. Three other negative arguments deserve consideration, however. These are (1) the argument by analogy with Ockham's treatment of grammatical features; (2) Spade's argument that signification "in the same way" requires syntactic identity in mental propositions¹¹; and (3) the argument from the fact that concepts bear a "natural likeness" to their objects.

4.1 The Analogy With Grammatical Features

A significant argument against the existence of mental synonymy derives from Ockham's criterion for

determining which grammatical features do and do not persist into mental language. As Spade puts it[17]:

... grammatical features of spoken or written language that do not serve the "needs of signification" by affecting truth conditions are not present in mental grammar. That this rules out all synonymy in mental language seems to be the clear intention of Ockham's whole discussion in Quodlibet 5, q. 8, with its repeated statement that what is in mental language is there only because of the "needs of signification", not for the sake of "decoration" or "well-formedness", and that synonymy does not serve the "needs of signification".

For example, such grammatical features as gender of nouns and the different conjugations of verbs do not persist into mental language, as they are irrelevant to signification - that is, they have no effect on truth conditions. Such features would be unnecessary in mental language. If they existed there, they would have the status of mere ornaments. It seems plausible that the differences between any synonymous expressions in written language would be equally unnecessary in mental language. Such differences would therefore not persist into mental language, implying that that mental language has no synonymy.

As an initial reply, one might point out that it is not unreasonable to suppose that there could be other relevant needs besides the "needs of signification." For example, I argued above that synonymous mental expressions might sometimes serve the needs of efficiency in mental language, by providing a simpler expression for complex thoughts. This would seem to be a more relevant reason for the presence of synonyms than mere "decoration." Grammatical features such as gender and different conjugations, of course, would not serve even this purpose (in fact they would achieve the opposite, by making mental expressions more complex than necessary), and so would not make it into mental language, but there are conceivably other instances of synonymy that would qualify.

As perhaps a more compelling reply, one might argue that certain instances of synonymy in mental language are not there because they serve any particular needs, but rather because they must exist as a byproduct of other properties of mental language. Logical synonymy would be one example of this. If we assume that mental language has a need for certain primitive logical operators, and for the ability to combine any expressions according to these operators, then as we saw above we are forced to the conclusion that some complex expressions must be synonymous. This fact in itself does not serve any particularly useful purpose for mental language; rather, it is a consequence of other facts about mental language.

This leads us to another line of reply to the argument above. We might argue that Ockham's criterion should apply to the determination of those grammatical features that persist into the *lexicon* - that is, the set of simple terms - of mental language, but that once we have determined the simple terms, then complex expressions should be derived directly from these according to combinatorial principles, just as English sentences are derived from words according to such principles. So although no two lexical items

in mental language might be synonymous, it could nevertheless be the case that certain complex expressions could be synonymous. To stipulate that two apparently quite different but synonymous complex expressions should in fact be identical in mental language would require ad hoc tinkering with the structure of mental language, as we saw above. By contrast, the elimination of irrelevant grammatical features from simple terms in mental language is not ad hoc at all, precisely because it can be achieved by a correct specification of the elements of mental language at the basic level.

Of course, if we accept this argument against lexical synonymy, it would follow that the argument from conceptual change, above, might have to be rejected. Although I am less convinced by the conceptual change argument than by the logical equivalence argument, I should nevertheless point out that a similar defense of it could be mounted. As with logical synonymy, the synonymy of two mental concepts that have drifted together is not in mental language to serve any need. Rather, it is there as a simple byproduct of the process of conceptual change, and of the way that conceptual change works. And the distinction between two synonymous terms might not be immediately eliminated from mental language, as irrelevant grammatical features are, because the synonymy might not be immediately apparent. If two independent concepts have by coincidence drifted together, there might be no obvious marker of their synonymy, so we might expect it to take a while before mental language became sensitive to the fact of their common meaning. The existence of synonymous mental terms would not serve any *purpose* in mental language, and we could expect that the terms might be merged at some future time, for instance when the subject becomes aware of their synonymy. Until that time, however, the synonymous terms might both exist, with their coexistence being a byproduct of the terms' divergent histories.

4.2 Signification "in the same way"

As we saw earlier, Ockham's criterion for the synonymy of two terms is that they not only signify the same things, but they signify those things *in the same way*. It is not clear what Ockham meant by signification "in the same way," but Spade has presented a construal of this phrase that, if correct, would immediately destroy any possibility of synonymy in mental language.[18] It is therefore vital that we come to grips with this interpretation of Ockham's criterion.

Spade's suggestion is that a "way" of signifying should be interpreted syntactically[19]:

A mental expression or concept signifies a thing x in a given syntactic mode m if and only if x is signified by some constituent non-complex categorematic term occurring within that mental expression in the grammatical or syntactical construction m .

If this is the case, then any two synonymous expressions must be made up of identical simple categorematic concepts, in identical syntactic constructions. In other words, they must be identical. It follows that there can be no synonymy in mental language.

The ease with which this argument buys its conclusion is suspicious. It seems to *define* mental synonymy out of existence; but we have already seen plenty of cases where the idea of mental synonymy at least

seems coherent. It may be for a variety of reasons that there turns out to be no synonymy in mental language, but this does not seem on the face of it to be an analytic truth. The case discussed earlier - that of synonymous, logically equivalent but distinct propositions - seems at least to be a possible example, even if certain facts about the way mental language functions might indicate that it is not actual. Therefore on the face of it it would seem unlikely that mental synonymy can be debarred definitionally.

Spade's argument for this definition is not entirely clear to me. He argues that a "way of signifying" should not be construed as one of Ockham's four "modes" of signification in S. log. I.33.20. He then leaps from this claim to the claim that a "way" should be construed purely syntactically, but the grounds for this transition are not obvious. The closest thing I can find to an argument for this conclusion is the following (in the context of a discussion of why 'blind' and 'sight' are not synonymous, despite both signifying sight)[21]:

Now, in the case of a nominal definition, or indeed of complex expressions generally, it is relatively clear what it might mean to signify a thing *x* "negatively". It could mean that the expression as a whole signifies *x* in virtue of some constituent non-complex categorematic term which signifies *x* and which occurs within the scope of a negation- sign in that expression. This is, in part at least, a syntactic criterion.

But it does not seem to me that "negative" signification must be construed purely syntactically. It does not seem unlikely that a semantic characterization could be arrived at. For instance, there might be a characterization that exploited the fact that although 'blind' and 'sight' signify the same *things*, expressions in which they are embedded have opposite truth-values (for instance, 'James is blind' is true precisely when 'James is sighted' is not true). This might well be more along the lines of what "signifying in the same way" comes to. Indeed, it is interesting to note that on the later "adverbial" theory of signification (used by Peter of Ailly, among others[22]), the same locution - that is, the locution of signify "in a given manner" - is used for that part of signification (of a proposition, in this case) from which truth-values derive. Perhaps this is not entirely coincidental. If we are to speculate, it does not seem unlikely that Ockham might have had a similar semantic criterion in mind, although it is quite possible that he never formalized the criterion, leaving it at the level of intuition.

Furthermore, even if we accept Spade's argument in the quoted passage, it seems to fall short of establishing that signification "in a given way" is a purely syntactic notion. All that is established is that "negative" signification - just a small aspect of what counts in determining a "way" of signifying - *might* have a criterion that is at least partly syntactic. This seems too weak to establish the broad conclusion suggested by Spade when he states,[23]: "The considerations of section VII, above, suggest that ... the 'ways' at stake here are ... *syntactic* modes of signification."

It seems to me that there is something wrong with the idea that there are syntactic criteria for a notion such as synonymy, which is a deeply semantic notion. It is more plausible that synonymy ought to be characterized purely semantically, in terms of the relationship between elements of language and their actual or possible referents. Looking at Ockham's criterion for synonymy, it seems to me that this is what

he may have been doing, although he left the crucial strengthening clause ("signify in the same way") somewhat vague. Nevertheless it seems intuitively plausible that there is an extra semantic criterion required of synonymous terms, over and mere coincidence of signification. This can be captured by the modern idea that this coextensiveness is necessary and *a priori*. Ockham may not have thought about the issue in explicitly this way, but the intuitive notion is clear enough. I would therefore suggest that this is what Ockham's strengthening clause in the criterion for synonymy should come to.

It seems to me that if we accept the claim that there is no synonymy in mental language, then Spade's syntactic construal of the strengthening criterion will be correct in practice - but this will be precisely because synonymy does not exist, so that a syntactic check will be sufficient to check for identity of meaning. It seems to me further that there are no compelling grounds, other than the nonexistence of synonymy, for supposing the syntactic construal to be correct. If so, then it follows that the syntactic construal cannot be used to provide support for the nonexistence of synonymy.

A further argument in favor of a semantic construal of the strengthening criterion is that it satisfies our intuition that mental expressions corresponding to 'man and (dog or cat)' and '(man and dog) or (man and cat)' are synonymous *whether or not* they are identical. For all these reasons, I believe that the case for the syntactic construal of "ways" of signification is at best inconclusive, and does not provide compelling evidence against the possibility of synonymy in mental language.

While on this topic, I should note that a semantic criterion for synonymy would answer another argument against mental synonymy, also given by Spade[24]:

What would equivocation or synonymy in mental language amount to? Since there is no supramental language to appeal to in the way one appeals to mental language to account for synonymy and equivocation in spoken and written language, how could it even arise in mental language?

This argument seems to assume that the only real criterion for mental synonymy could be appeal to some higher language²⁵ - that is, the criterion must be a *formal* criterion. It seems to ignore the possibility of a semantic criterion, in terms for instance of referents and/or truth-values. If we have a semantic criterion for synonymy, the problem posed here by Spade is no problem at all.

4.3 Natural Likenesses

The final argument rests on Ockham's contention that concepts bear natural likenesses to their objects. Spade argues that it is unlikely that two concepts could bear a natural likeness in the relevant sense to their objects, but still be more than numerically distinct.[26] If this is correct, it would follow that mental synonyms could not exist.

The reply to this argument is simply to say that it does not seem too implausible that distinct mental terms could bear natural likenesses to a common object, especially if they are complex expressions rather than simple terms. Going back to our favorite example, it seems plausible that the two concepts 'man and

(dog or cat)' and '(man and dog) or (man and cat)' both bear a natural likeness to their object, despite the apparent difference in the expressions. These two mental expressions seems to differ in form or structure, but they are nevertheless concepts of exactly the same things. It might be difficult to imagine how two distinct *simple* terms could bear the relevant likenesses simultaneously, but this difficulty vanishes in the case of complex expressions. So this does not seem to be a compelling objection to mental synonymy.

5 Simple Connotative Terms

Before concluding, I should note the possible effect of the discussion above on the question of whether there exist simple connotative terms in mental language. Spade has argued that simple connotative mental terms cannot exist, as if they did, they would be synonymous with their expanded nominal definitions[27]; but of course there is no synonymy in mental language! Panaccio has pointed out another possibility: that simple connotative terms exist, but in fact are not synonymous with their nominal definitions.[28]

The present discussion raises a third possibility: that simple connotative terms exist, and are synonymous with their nominal definitions. Spade and Panaccio reject this possibility because of the supposed impossibility of mental synonymy; but if we accept the above arguments for the possibility of mental synonymy, then this becomes a live option. Indeed, once we have gotten over the hurdle of mental synonymy, it may even fit certain textual evidence better than either of these claims, as Ockham certainly claimed that simple connotative terms exist, and there are strong reasons to believe that connotative terms are synonymous with their definitions (although Panaccio's argument may have weakened this evidence somewhat).

The discussion I have given here even provides a reason why simple connotative terms might exist: namely, to act as abbreviations for their nominal definitions, for the sake of making mental language less cumbersome. This is perhaps the reason why many connotative terms exist in everyday languages. It does not seem wholly unreasonable that they might exist in mental language for the same purpose.

Of course, I do necessarily wish to maintain that Ockham actually held this position, as the existence of mental synonymy seems to contradict textual evidence, at least as the evidence has traditionally been interpreted. Nevertheless I should point out that this instance of mental synonymy is less difficult to reconcile with textual evidence than the other instances we have looked at. This is because we can maintain the synonymy between simple connotative terms and their nominal definitions in mental language without giving up Ockham's claim that spoken or written synonyms are subordinated to identical mental expressions. Instead, we can hold that a spoken or written connotative term and its nominal definition are both subordinated to the corresponding term (the complex expression) in mental language. The simple connotative mental term is introduced entirely within the mental sphere, for the purposes of abbreviation. No spoken or written term is subordinated to it, as it is solely a construct introduced for mental efficiency, just as connotative terms in spoken or written language are frequently introduced for efficiency.

While this position may or may not be independently plausible, we should note that it seems to be at least as compatible with the textual evidence as Spade's and Panaccio's accounts. The only claim of Ockham that it might explicitly contradict is the passage from Quodlibet V.8,[29] and even that passage might be interpreted in a manner compatible with the position maintained here. In any case, whether or not this position was actually held by Ockham, it seems quite attractive as a theoretical possibility.

6 Conclusion

I have done my best to act as an advocate for the possibility that synonymy could exist in mental language. Now is the time to pass considered judgment.

First, the question: did Ockham believe there is synonymy in mental language? To this I believe the answer is probably no. For a start, he seems to deny it on one occasion. Furthermore, as we have seen, two of my arguments for synonymy have required overturning a claim that Ockham made a number of times, that is, the claim that synonyms in spoken or written language are subordinated to identical terms in mental language. On the other hand, the third argument, the argument from efficiency, is independent of this claim, as we can suppose that abbreviation of complex expressions could be entirely internal to the mental system; and this argument also gains some additional textual support from Ockham's otherwise anomalous claim that simple connotative terms exist in mental language. Still, overall, it seems most plausible to me that Ockham was at least implicitly committed to the nonexistence of mental synonyms, although we have seen that internal tensions arise in his theory due to this commitment.

Next, the question: should Ockham have believed that there is synonymy in mental language? To this, my answer is less clear. I believe that there is a very strong argument for the possibility of mental synonymy, in the argument from logical equivalence. As far as I can tell, Ockham never considered this argument. If he had, it is possible that it might have caused him at least to have some doubts about the impossibility of synonymy. Even if he had eventually come down against the possibility, he might have been required to modify his theory in some serious way, or at least to make a number of its features more explicit. As for the second and third argument, those from conceptual change and from efficiency, I believe that neither of these is conclusive, but that both provide some added plausibility for the notion of mental synonymy, by demonstrating how it might come about and even perhaps serve a useful purpose. Finally, Ockham's belief that synonymy in spoken or written language can go unrecognized seems to yield the existence of synonymous mental terms as a natural consequence. This provides further grounds for thinking that whether or not Ockham accepted synonymy, he ought to have.

Of the negative arguments (apart from the textual considerations), I believe that the argument from analogy with grammatical features has been cast into doubt, by showing how certain instances of synonymy might exist without having to serve any need, but rather because they are byproducts of other features of mental language. The argument from signifying "in the same way" seems to me to be quite inconclusive, due to the difficulty in ascertaining just what Ockham meant by his phrase. It does not seem too unlikely that he had a semantic criterion in mind, rather than a syntactic criterion. It is possible that the final argument, from natural likenesses, could be made into a strong argument against mental

synonymy with some work. I have been concerned to reply to the argument as it is given by Spade, but it is not impossible that stronger versions of the argument exist. Still, such arguments would have to deal with the apparent existence of complex expressions that are distinct but clearly have the same meaning. If some version of natural likeness theory had the consequence that these expressions are not synonymous, then it might not be unreasonable to suggest that that version of the theory should be thrown out, rather than the possibility of synonymy.

Overall, I must adjudicate the case "not proven," but with some strong evidence in favor of mental synonymy that must be dealt with before a retrial. Whichever way the verdict comes down, it seems to me that the possibility of mental synonymy is not as objectionable as has commonly been supposed.

Notes

0 This paper was written when I was a graduate student in Paul Spade's class on Medieval Logic at Indiana University in 1991. I owe an enormous debt to Spade for his insights and encouragement.

1 Trentman 1970; Spade 1980.

2 S. log. I.6 (OPh I.19.8-14).

3 Spade 1975.

4 To be strictly accurate, it is even weaker than this, as for instance a term and its negation are taken to signify the same thing but are not intersubstitutable. DELETE

5 Contrary to what Spade 1996, 109, seems to suggest.

6 S. log. I.6 (OPh I.19.8-13; Ockham 1995b, 16-17).

7 And probably is not so preserved, if one accepts the arguments of Gregory of Rimini and Peter of Ailly. See Spade 1996, 120-127.

8 Quodl. V.9 (OTh IX.513.11-12). Compare Quodl. V.8 concl. 2 (OTh IX.510-511.65-69).

9 For example, expressions corresponding to connotative terms, on Spade's account (Spade 1975).

10 See ?? below.

11 Spade 1975.

12 OTh IX.513.130-136; Spade 1996, 358-359.

13 Spade 1980, 12.

14 Ibid.

15 See II.1 and II.2 above.

16 S. log. I.6 (OPh I.19.8-13; Ockham 1995b, 16-17).

17 Spade 1980, 12.

18 Spade 1975 and 1980.

19 Spade 1975, 68.

20 OPh I.95-96.

21 Spade 1975, 67.

22 See Spade 1996, 180-182.

23 Spade 1975, 68.

24 Spade 1996, 99.

25 Spade acknowledges this, *ibid.*

26 Spade 1996, 99-100.

27 Spade 1975.

28 Panaccio 1990.

29 See III.1 above.

Moving Forward on the Problem of Consciousness

David J. Chalmers

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

This paper is a response to the commentaries in the *Journal of Consciousness Studies* on my paper "[Facing Up to the Problem of Consciousness](#)." I have written it so that it can be understood independently of the commentaries, however, and so that it provides a detailed elaboration and extension of some of the ideas in the original paper.

The 26 commentaries (with links to online versions when they exist) were by [Bernard Baars](#), Douglas Bilodeau, [Patricia Churchland](#), [Tom Clark](#), C.J.S. Clarke, [Francis Crick & Christof Koch](#), [Daniel Dennett](#), [Stuart Hameroff & Roger Penrose](#), Valerie Hardcastle, David Hodgson, [Piet Hut & Roger Shepard](#), Benjamin Libet, E.J. Lowe, [Bruce MacLennan](#), Colin McGinn, Eugene Mills, Kieron O'Hara & Tom Scutt, Mark Price, William Robinson, Gregg Rosenberg, William Seager, Jonathan Shear, Henry Stapp, [Francisco Varela](#), [Max Velmans](#), and Richard Warner. The complete symposium has been collected into a book, [Explaining Consciousness: The Hard Problem](#), published by MIT Press.

CONTENTS

- [Introduction](#)
- [Deflationary critiques](#)
 - [Deflationary analogies](#) (Dennett, Churchland)
 - [Is explaining the functions enough?](#) (Dennett, Churchland)
 - [Type-B materialism](#) (Clark, Hardcastle)
 - [Other deflationary approaches](#) (McGinn, O'Hara & Scutt, Price)
- [Nonreductive analyses](#)
 - [Conceptual foundations](#) (Hodgson, Lowe, Velmans)
 - [The roots of the hard problem](#) (McGinn, Robinson, Warner)
 - [Fundamental laws](#) (Mills, Price)
 - [Epiphenomenalism and interactionism](#) (Hodgson, Seager, Stapp, Warner)
 - [My psychophysical laws](#) (Hardcastle, Libet, Lowe, Seager, Varela, Velmans)
- [Positive proposals](#)
 - [Neuroscientific and cognitive approaches](#) (Baars, Crick & Koch, MacLennan)
 - [Phenomenological approaches](#) (Shear, Varela)
 - [Physics-based approaches](#) (Bilodeau, Clarke, Hameroff & Penrose, McGinn, Stapp)
 - [Fundamental psychophysical theories](#) (Hut & Shepard, Rosenberg, Seager)

- [Conclusion](#)

1 INTRODUCTION

I am very grateful to all the contributors to this symposium for their thoughtful comments. The various papers reflect a wide range of approaches and of views, yielding a rich snapshot of the current state of play on the problem of consciousness. There are some interesting criticisms of my point of view, which I hope to address in this reply in a way that clarifies the central issues at hand, and there are also a number of intriguing positive proposals for confronting the problem. I am honored to have provided an opportunity to bring such a thought-provoking collection of ideas together.

When I wrote my paper, I had no idea that it would be subject to such close analysis. That may be a good thing, as all the hedges, qualifications, and citations I would have added if I had known might have made the paper close to unreadable, or at any rate twice the size. But it also means that the paper - intended as a crisp presentation of some central issues, mostly for non-philosophers - skates quickly over some subtleties and has less flesh on its bones than it might. I will try to flesh out the picture in this piece, while still keeping the discussion at a non-technical level. A more detailed presentation can be found in my book *The Conscious Mind*, to which I will occasionally point in this response.

Because of the unexpected influence of the "hard problem" formulation, I have occasionally received far more credit than I deserve. So let me state the obvious: the reason the formulation has caught on is that everyone knew what the hard problem was all along. The label just makes it a little harder to avoid. Any number of thinkers in the recent and distant past - including a number of contributors to this symposium - have recognized the particular difficulties of explaining consciousness and have tried to face up to them in various ways. All my paper really contributes is a catchy name, a minor reformulation of philosophically familiar points, and a specific approach to dealing with them.

The papers in the symposium are divided fairly evenly into those that take issue with aspects of my analysis, and those that provide positive approaches of their own. I will concentrate mostly on those in the first class, though I will make a few comments on those in the second. A quick glance at the relevant papers may give the appearance of much disagreement and a sprawling landscape of mutually contradictory points of views; but I think a closer look reveals a much more coherent picture. Once a few minor misunderstandings and verbal disagreements are cleared up, and the various contributions are aligned, one is left with a small number of central "choice points" on which the central disagreements turn. I hope that my reply helps to clarify this landscape.

The reply has three main parts. In the first I consider the critiques of a generally reductive or "deflationary" orientation; in the second I consider those of a generally nonreductive orientation; and in the third I make some comments on the various positive proposals.

2 DEFLATIONARY CRITIQUES

Recall the main conceptual distinction between the easy and hard problems. The easy problems - explaining discrimination, integration, accessibility, internal monitoring, reportability, and so on - all concern the performance of various *functions*. For these phenomena, once we have explained how the relevant functions are performed, we have explained what needs to be explained. The hard problem, by contrast, is not a problem about how functions are performed. For any given function that we explain, it remains a nontrivial further question: why is the performance of this function associated with conscious experience? The sort of functional explanation that is suited to answering the easy problems is therefore not automatically suited to answering the hard problem.

There are two quite different ways in which a materialist might respond to this challenge. The *type-A* materialist denies that there is a "hard problem" distinct from the "easy" problems; the *type-B* materialist accepts (explicitly or implicitly) that there is a distinct problem, but argues that it can be accommodated within a materialist framework all the same. Both of these strategies are taken by contributors to this symposium. I will discuss the first strategy in the next two sections, and the second strategy after that.

2.1 Deflationary analogies

The type-A materialist, more precisely, denies that there is any phenomenon that needs explaining, over and above explaining the various functions: once we have explained how the functions are performed, we have thereby explained everything. Sometimes type-A materialism is expressed by denying that consciousness exists; more often, it is expressed by claiming that consciousness may exist, but only if the term "consciousness" is defined as something like "reportability", or some other functional capacity. Either way, it is asserted that there is no interesting fact about the mind, conceptually distinct from the functional facts, that needs to be accommodated in our theories. Once we have explained how the functions are performed, that is that.

Note that type-A materialism is not merely the view that consciousness is identical to some function, or that it plays a functional role, or that explaining the functions will help us explain consciousness. It is the much stronger view that there is not even a distinct *question* of consciousness: once we know about the functions that a system performs, we thereby know everything interesting there is to know. Type-A materialism subsumes philosophical positions such as eliminativism, behaviorism, analytic functionalism, and others, but it does not include positions (such as those embraced by **Clark** and **Hardcastle**) that rely on an *a posteriori* identity between consciousness and some physical/functional property. Positions of the latter sort accept that there is a real phenomenon to be accounted for, *conceptually* distinct from the performance of functions (the *a posteriori* identity ties together *a priori* distinct concepts), and therefore count as type-B materialism. Type-A materialism, by contrast, denies that there is a conceptually distinct explanatory target at all.

This is an extremely counterintuitive position. At first glance, it seems to simply deny a manifest fact about us. But it deserves to be taken seriously: after all, counterintuitive theories are not unknown in science and philosophy. On the other hand, to establish a counterintuitive position, strong arguments are needed. And to establish *this* position - that there is really nothing else to explain - one might think that extraordinarily strong arguments are needed. So what arguments do its proponents provide?

Perhaps the most common strategy for a type-A materialist is to deflate the "hard problem" by using analogies to other domains, where talk of such a problem would be misguided. Thus **Dennett** imagines a vitalist arguing about the hard problem of "life", or a neuroscientist arguing about the hard problem of "perception". Similarly, **Paul Churchland** (1996) imagines a nineteenth century philosopher worrying about the hard problem of "light", and **Patricia Churchland** brings up an analogy involving "heat". In all these cases, we are to suppose, someone might once have thought that more needed explaining than structure and function; but in each case, science has proved them wrong. So perhaps the argument about consciousness is no better.

This sort of argument cannot bear much weight, however. Pointing out that analogous arguments do not work in other domains is no news: the whole point of anti-reductionist arguments about consciousness is that there is a disanalogy between the problem of consciousness and problems in other domains. As for the claim that analogous arguments in such domains might once have been plausible, this strikes me as something of a convenient myth: in the other domains, it is more or less *obvious* that structure and function are what need explaining, at least once any experiential aspects are left aside, and one would be hard pressed to find a substantial body of people who ever argued otherwise.

When it comes to the problem of life, for example, it is just obvious that what needs explaining is structure and function: How does a living system self-organize? How does it adapt to its environment? How does it reproduce? Even the vitalists recognized this central point: their driving question was always "How could a mere physical system perform these complex functions?", not "Why are these functions accompanied by life?" It is no accident that Dennett's version of a vitalist is "imaginary". There is no distinct "hard problem" of life, and there never was one, even for vitalists.

In general, when faced with the challenge "explain X", we need to ask: what are the phenomena in the vicinity of X that need explaining, and how might we explain them? In the case of life, what cries out for explanation are such phenomena as reproduction, adaptation, metabolism, self-sustenance, and so on: all complex functions. There is not even a plausible candidate for a further sort of property of life that needs explaining (leaving aside consciousness itself), and indeed there never was. In the case of consciousness, on the other hand, the manifest phenomena that need explaining are such things as discrimination, reportability, integration (the functions), *and experience*. So this analogy does not even get off the ground.

Or take **Churchland's** example of heat. Here, what cries out for explanation are such things as: heat's abilities to expand metals, the causation of fire, heat transmission between substances, the experience of hotness. All but the last of these are clearly functions, and it is these functions that reductive explanations of heat explain. The existence of such functions is entailed by the microphysical story about heat: in any world that is physically identical to ours, such functions will automatically be present.

If someone were to claim that something were "left out" by reductive explanations of heat (as Churchland suggests they might), or of light (as Paul Churchland suggests they might), what something might they be referring to? The only phenomenon for which the suggestion would be even remotely plausible is our

subjective experience of light and of hotness. The molecular theory of heat does not explain the sensation of heat; and the electromagnetic theory of light does not explain what it is like to see. And understandably so: the physicists explaining heat and light have quite reasonably deferred the explanation of their experiential manifestations until the time when we have a reasonable theory of consciousness. One need not explain everything at once. But with consciousness itself, subjective experience is precisely what is at issue, so we cannot defer the question in the same way. Thus once again, the analogy is no help to a reductionist.

In his article "The Rediscovery of Light" (1996), **Paul Churchland** suggests that parallel antireductionist arguments could have been constructed for the phenomenon of "luminescence", and might have been found plausible at the time. I have my doubts about that plausibility, but in any case it is striking that his arguments about luminescence all depend on intuitions about the conscious experience of light. His hypothetical advocate of a "hard problem" about light appeals to light's "visibility" and the "visual point of view"; his advocate of a "knowledge argument" about light appeals to blind Mary who has never had the experience of seeing; and the advocate of a "zombie" argument appeals to the conceivability of a universe physically just like ours, but in which everything is dark. That the first two arguments trade on intuitions about experience is obvious; and even for the third, it is clear on a moment's reflection that the only way such a universe might make sense is as a universe in which the same electromagnetic transmission goes on, but in which no-one has the experience of seeing.

Churchland might insist that by "luminescence" he means something quite independent of experience, which physical accounts still do not explain: but then the obvious reply is that there is no good reason to believe in luminescence in the first place. Light's structural, functional, and experiential manifestations exhaust the phenomena that cry out for explanation, and the phenomena in which we have any reason to believe. By contrast, conscious experience presents itself as a phenomenon to be explained, and cannot be eliminated in the same way.

A similar critique applies to such examples such as Dennett's "cuteness" (what needs explaining is the structure and functioning of cute people, and our experience and judgment of them as cute), his "perception" (the functioning of perceptual systems plus the experience of perception), and so on. In all such cases, either the analogous arguments are not even *prima facie* plausible (as in the case of life), or at best, they gain their plausibility through pointing to experiential properties that reductive accounts omit (as in the cases of perception and light). So they can do no work at all in arguing for reductionism about experience.

Indeed, similar remarks can be made about *any* phenomenon that we observe in the external world. When we observe external objects, we observe their structure and function; that's all. Such observations give no reason to postulate any new class of properties, except insofar as they explain structure and function; so there can be no analog of a "hard problem" here. Even if further properties of these objects existed, we could have no access to them, as our external access is physically mediated: such properties would lie on the other side of an unbridgeable epistemic divide. Consciousness uniquely escapes these arguments by lying at the center of our epistemic universe, rather than at a distance. In this case alone, we can have access to something other than structure and function.

2.2 Is explaining the functions enough?

So, analogies don't help. To have any chance of making the case, a type-A materialist needs to *argue* that for consciousness, as for life, the functions are all that need explaining. Perhaps some strong, subtle, and substantive argument can be given, establishing that once we have explained the functions, we have automatically explained everything. If a sound argument could be given for this surprising conclusion, it would provide as valid a resolution of the hard problem as any.

Is there any compelling, non-question-begging argument for this conclusion? The key word, of course, is "non-question-begging". Often, a proponent will simply *assert* that functions are all that need explaining, or will argue in a way that subtly *assumes* this position at some point. But that is clearly unsatisfactory. Prima facie, there is very good reason to believe that the phenomena a theory of consciousness must account for include not just discrimination, integration, report, and such functions, but also *experience*, and prima facie, there is good reason to believe that the question of explaining experience is distinct from the questions about explaining the various functions. Such prima facie intuitions can be overturned, but to do so requires very solid and substantial argument. Otherwise, the problem is being "resolved" simply by placing one's head in the sand.

Upon examining the materialist papers in this symposium, such arguments are surprisingly hard to find. Indeed, despite their use of various analogies, very few of the contributors seem willing to come right out and say that in the case of consciousness, the functions are all that need explaining. Only **Dennett** embraces this position explicitly, and even he does not spend much time *arguing* for it. But he does spend about a paragraph making the case: presumably this paragraph bears the weight of his piece, once the trimmings are stripped away. So it is this paragraph that we should examine.

Dennett's argument here, interestingly enough, is an appeal to phenomenology. He examines his own phenomenology, and tells us that he finds nothing other than functions that need explaining. The manifest phenomena that need explaining are his *reactions* and his *abilities*; nothing else even presents itself as needing to be explained.

This is daringly close to a simple denial - one is tempted to agree that it might be a good account of *Dennett's* phenomenology - and it raises immediate questions. For a start, it is far from obvious that even all the items on Dennett's list - "feelings of foreboding", "fantasies", "delight and dismay" - are purely functional matters. To assert without argument that all that needs to be explained about such things are the associated functions seems to beg the crucial question at issue. And if we leave these controversial cases aside, Dennett's list seems to be a systematically incomplete list of what needs to be explained in explaining consciousness. One's "ability to be moved to tears" and "blithe disregard of perceptual details" are striking phenomena, but they are far from the most obvious phenomena that I (at least) find when I introspect. Much more obvious are the experience of emotion and the phenomenal visual field themselves; and nothing Dennett says gives us reason to believe that these do not need to be explained, or that explaining the associated functions will explain them.

What might be going on here? Perhaps the key lies in what Dennett has elsewhere described as the foundation of his philosophy: "third-person absolutism". If one takes the *third-person* perspective on oneself -- viewing oneself from the outside, so to speak - these reactions and abilities are no doubt the main focus of what one sees. But the hard problem is about explaining the view from the *first-person* perspective. So to shift perspectives like this - even to shift to a third-person perspective on one's first-person perspective, which is one of Dennett's favorite moves - is again to *assume* that what needs explaining are such functional matters as reactions and reports, and so is again to argue in a circle.

Dennett suggests "subtract the functions and nothing is left". Again, I can see no reason to accept this, but in any case the argument seems to have the wrong form. An analogy suggested by Gregg Rosenberg is useful here. Color has properties of hue, saturation, and brightness. It is plausible that if one "subtracts" hue from a color, nothing phenomenologically significant is left, but this certainly doesn't imply that color is nothing but hue. So even if Dennett could argue that function was somehow *required* for experience (in the same way that hue is required for color), this would fall a long way short of showing that function is all that has to be explained.

A slight flavor of non-circular argument is hinted at by Dennett's suggestion: "I wouldn't know what I was thinking about if I couldn't identify them by their functional differentia". This tantalizing sentence suggests various reconstructions, but all the reconstructions that I can find fall short of making the case. If the idea is that functional role is essential to the (subpersonal) *process* of identification, this falls short of establishing that functioning is essential to the experiences themselves, let alone that functioning is all there *is* to the experiences. If the idea is rather than function is all we have access to at the *personal* level, this seems false, and seems to beg the question against the intuitive view that we have knowledge of intrinsic features of experience. But if Dennett can elaborate this into a substantial argument, that would be a very useful service.

In his paper, Dennett challenges me to provide "independent" evidence (presumably behavioral or functional evidence) for the "postulation" of experience. But this is to miss the point: conscious experience is not "postulated" to explain other phenomena in turn; rather, it is a phenomenon to be explained in its own right. And if it turns out that it cannot be explained in terms of more basic entities, then it must be taken as irreducible, just as happens with such categories as space and time. Again, Dennett's "challenge" *presupposes* that the only explananda that count are functions.[*]

*[[[Tangentially: I would be interested to see Dennett's version of the "independent" evidence that leads physicists to "introduce" the fundamental categories of space and time. It seems to me that the relevant evidence is spatiotemporal through and through, just as the evidence for experience is experiential through and through.]]]

Dennett might respond that I, equally, do not give *arguments* for the position that something more than functions needs to be explained. And there would be some justice here: while I do argue at length for my conclusions, all these arguments take the existence of consciousness for granted, where the relevant concept of consciousness is explicitly distinguished from functional concepts such as discrimination, integration, reaction, and report. Dennett presumably disputes this starting point: he thinks that the only sense in which people are conscious is a sense in which consciousness is *defined* as reportability, as a

reactive disposition, or as some other functional concept.

But let us be clear on the dialectic. It is *prima facie* obvious to most people that there is a further phenomenon here: in informal surveys, the large majority of respondents (even at Tufts!) indicate that they think something more than functions needs explaining. Dennett himself - faced with the results of such a survey, perhaps intending to deflate it - has accepted that there is at least a *prima facie* case that something more than functions need to be explained; and he has often stated how "radical" and "counterintuitive" his position is. So it is clear that the default assumption is that there is a further problem of explanation; to establish otherwise requires significant and substantial argument.

I would welcome such arguments, in the ongoing attempt to clarify the lay of the land. The challenge for those such as Dennett is to make the nature of these arguments truly clear. I do not think it a worthless project - the hard problem is so hard that we should welcome all attempts at a resolution - but it is clear that anyone trying to make such an argument is facing an uphill battle.[*]

*[[[One might look to Dennett's book *Consciousness Explained* for non-circular arguments, but even here such arguments for the relevant conclusion are hard to find. The plausible attacks on a "place in a brain where it all comes together" do nothing to remove the hard problem. The book's reliance on "heterophenomenology" (verbal reports) as the central source of data occasionally slips into an unargued assumption that such reports are all that need explaining, especially in the discussion of "real seeming", which in effect assumes that the only "seemings" that need explaining are dispositions to react and report. I think there may be a substantial argument implicit in the "Orwell/Stalin" discussion - essentially taking *materialism* as a premise and arguing that if materialism is true then the functional facts exhaust all the facts - but even this is equivalent to "if something more than functions needs explaining, then materialism cannot explain it", and I would not disagree. At best, Dennett's arguments rule out a middle-ground "Cartesian materialism"; the hard problem remains as hard as ever.]]]

In **Churchland's** paper, this sort of argument is even harder to find. Indeed, it is not always clear who Churchland is arguing with: she does not address the central arguments in the keynote paper at any point, and she often seems to be arguing with someone with views quite different from mine. Her arguments have *premises* that are consistently more plausible than Dennett's, but they do not come close to establishing the relevant conclusion. I include Churchland as a type-A materialist as she suggests that there is no principled difference between the "hard" and "easy" problems, but her position is sufficiently inexplicit that it is hard to know for sure.

Churchland asks for a systematic difference between the "easy" and "hard" problems, not mentioning the detailed analysis of this difference in my paper. The difference is, of course, that the easy problems are all clearly problems of explaining how functions are performed, and the hard problem is not. Perhaps Churchland, like Dennett, would deny this; unlike Dennett, however, she never addresses the question directly. If she truly holds that the functions (discrimination, integration, access, control, report, ...) are all that we need to account for, then clearly some explicit argument is required. If she does not, then the relevant distinction is present right there.

Churchland notes correctly that phenomena such as attention have an experiential component. I am not sure how this is meant to deflate the problem of experience. Vision has an experiential component, too;

that's the "hard" part. We can give neural or cognitive accounts of the functions associated with these phenomena, but it remains unclear why the experiential aspect should accompany those functions. This isn't to deny that it *does* accompany them. There are deep and intimate links between the "hard" and "easy" phenomena, of which I note some in my paper, and more in my book. So when Churchland criticizes somebody's proposal for ruling out such links, it is not my proposal she is addressing.

Perhaps the problem is that Churchland sets up the "easy"/"hard" distinction as the distinction between the problems of (e.g.) attention, learning, and short-term memory on one hand, and the problem of consciousness on the other. This is not quite my way of doing things: I set up the distinction as that between explaining how functions are performed and explaining subjective experience. It is plausible that the notions of "memory", "attention", and perhaps even "consciousness" subsume elements both of functioning and of subjective experience, as Churchland in effect points out - so there are "easy" and "hard" aspects of memory, attention, and consciousness. To keep things clear, it is best to set up the distinction directly.

Churchland is also right to note that it is not always obvious just where experience is present and where it is not, especially in fringe cases. But it is a philosophical truism that we should not let the existence of fringe cases blind us to the facts about clear cases. One goal of a theory of experience will be to clarify the status of those fringe cases; in the meantime, in cases where experience is clearly present, it is as hard to explain as ever.

And Churchland is also quite right that there is much about the "easy" problems that we do not understand. "Easy" is of course a term of art, and nothing substantial in my arguments rests on it. Churchland's point would be a relevant rebuttal to an argument that rested on it, or to an argument from ignorance, but my argument is nothing of the sort. Facts of the form "we don't know" or "I can't imagine" play no explicit or implicit role in my arguments. Rather, the key is the conceptual point: the problem of consciousness is not a problem about how functions are performed. No matter how much we find out about the mechanisms that perform these functions, the basic explanatory point is unaffected.

Contrast Churchland's case of sensorimotor integration. It's true that we do not know much about the mechanisms here. But we do know what we need to do to explain sensorimotor integration: we need to explain how information from different sensory areas is brought together and put to use in the control of action. This is a problem about how functions are performed: it is *guaranteed* that once we find the mechanism that performs the function and explain how it works, we will have explained sensorimotor integration. But for consciousness, this guarantee fails: it is not just functions that need to be explained. So the research program that promises so much on the easy problems needs to be augmented where the hard problem is concerned.

So Churchland either needs to *argue* that functions are all that need to be explained, or she needs to face up to the disanalogy and the explanatory problem directly. Homilies about the progress of science do not carry much weight in this context. We have seen that "normal" (function-explaining) science in the neuroscientific mode has limitations that have to be confronted, not ignored; and if one relies instead on a gesture in the direction of a major conceptual revolution sometime in the future, then one is in effect

conceding that the hard problem is very hard indeed.

Proponents of the "no problem" view sometimes like to suggest that their view is supported by the results of modern science, but all the science that I know is quite neutral here: I have never seen any experimental result that implies that functions are all that need to be explained. Rather, this view seems to be rooted in a *philosophical* claim. This claim does not seem to be supported either by empirical evidence or by non-circular argument; at the end of the day, it may be that the position is grounded instead in some sort of unargued axiom, such as Dennett's third-person absolutism. And to anyone who is impressed by the first-person phenomenology of consciousness, such an axiom will always beg the crucial questions. The position reduces to an unargued denial.

This is not to say that type-A materialism cannot be argued for at all. There are a few sophisticated arguments for such a position in the literature (for example, Shoemaker 1975 and White 1986), but even these ultimately come down to "consider the alternatives", isolating the difficulties that one gets into if one accepts that there is a further phenomenon that needs explaining. There is no doubt that these difficulties (both ontological and epistemological) are considerable; life would be a lot easier if the hard problem did not exist. But I think these difficulties are solvable; and in any case, to deny the problem because of the difficulties has the flavor of solution by decree. So while I think such arguments need to be taken very seriously, they do little to actually remove the problem. To truly make the problem go away, one needs *positive* and *non-circular* arguments for the counterintuitive conclusion that the functions are all that need explaining; and such arguments are very hard to find.

Of course, type-A materialism is unlikely to disappear any time soon, and we will probably just have to get used to the fact that there is a basic division in the field: that between those who think the "easy" problems are the only problems, and those who think that subjective experience needs to be explained as well. We can therefore expect two quite distinct sorts of theories of consciousness: those which explain the functions and then say "that's all", and those which take on an extra burden. In the end, the most progress will probably come from internal advances in the respective research programs, rather from the endless battle between the two. So beyond a certain point in the argument, theorists in these camps might just agree to disagree and get on with their respective projects. This way, everyone can move forward.

2.3 Type-B materialism

Type-A materialism offers a clean and consistent way to be a materialist, but the cost is that it seems not to take consciousness seriously. Type-B materialism tries to get the best of both worlds. The type-B materialist accepts that there is a phenomenon that needs to be accounted for, conceptually distinct from the performance of functions, but holds that the phenomenon can still be explained within a materialist framework. This is surely the most attractive position at a first glance. It promises to avoid the extremes of both hard-line reductionism and property dualism, which respectively threaten to deny the phenomenon and to radically expand our ontology.

I was attracted to type-B materialism for many years myself, until I came to the conclusion that it simply

cannot work. The basic reason for this is simple. Physical theories are ultimately specified in terms of structure and dynamics: they are cast in terms of basic physical structures, and principles specifying how these structures change over time. Structure and dynamics at a low level can combine in all sort of interesting ways to explain the structure and function of high-level systems; but still, structure and function only ever adds up to more structure and function. In most domains, this is quite enough, as we have seen, as structure and function are all that need to be explained. But when it comes to consciousness, something other than structure and function needs to be accounted for. To get there, an explanation needs a further ingredient.

The type-A materialist gets around this problem by asserting that for consciousness, too, structure and function are all that need to be explained. But this route is not open to the type-B materialist. Given that we have accepted that something more than structure and function needs to be accounted for, we are forced to the conclusion that the "further question" will arise for *any* account of physical processing: why is this structure and function accompanied by conscious experience? To answer this question, we need to supplement our story about structure and function with something else; and in doing so we move beyond truly reductive explanation.

So while many people think they can reject a Dennett-style "no problem" view and still expect a purely physical explanation of consciousness one day, this view seems untenable for systematic reasons. An account of physical processing may provide the *bulk* of a theory of human consciousness; but whatever account of processing we give, the vital step - the step where we move from facts about structure and function to facts about experience - will always be an *extra* step, requiring some substantial principle to bridge the gap. To justify this step, we need a new component in our theories.

There is one route to type-B materialism that one might think remains open; this is the route taken by **Clark** and **Hardcastle**. These two are clearly realists about phenomenal consciousness, and they are equally clearly materialists. They reconcile the two by embracing an empirical *identity* between conscious experiences and physical processes. Although consciousness is not equivalent *a priori* to a structural or functional property (as type-A materialists might suggest), the two are nevertheless identical *a posteriori*. We establish this identity through a series of *correlations*: once we find that consciousness and certain physical processes are correlated, the best hypothesis is that the two are identical. And this postulated identity bridges the explanatory gap.

This is a popular approach, but it has a number of problems. The problems are all rooted in the same place: it makes the identity an *explanatorily primitive* fact about the world. That is, the fact that certain physical/functional states are conscious states is taken as a brute fact about nature, not itself to be further explained. But the only such explanatorily primitive relationships found elsewhere in nature are fundamental laws; indeed, one might argue that this bruteness is precisely the mark of a fundamental law. In postulating an explanatorily primitive "identity", one is trying to get something for nothing: all of the explanatory work of a fundamental law, at none of the ontological cost. We should be suspicious of such free lunches; and indeed, I think there is something deeply wrong with the idea.

To evaluate the truth of materialism, what matters is whether all facts follow from the physical facts. As I

argue at length in my book, in *most* domains it seems that they certainly do. The low-level facts about physical entities determine the facts about physical structure and function at all levels with conceptual necessity, which is enough to determine the facts about chemistry, about biology, and so on. The facts about genes "fall out" of the facts about the structure and function of DNA, for example. A geneticist does not need a primitive Genetic Identity Hypothesis to cross the divide - "what do you know, whenever there is some unit that encodes and transmits hereditary characteristics, there is a gene!". Rather, to encode and transmit such characteristics is roughly all it *means* to be a gene; so there is an *a priori* implication from the facts about the structure and functioning of DNA in a reproductive context to the facts about genes. Even Mary in her black-and-white room could figure out the facts about genes, in principle, if she was equipped with the facts about DNA and the concepts involved.

But the facts about consciousness do not just fall out of the facts about the structure and functioning of neural processes, at least once type-A materialism is rejected. As usual, there is a further question -- "why are these processes accompanied by consciousness?" - and merely repeating the story about the physical processes does not provide an answer. If we have rejected type-A materialism, there can be no conceptual implication from one to the other.

Clark and Hardcastle's answer is to *augment* one's account of physical processes with an "identity hypothesis" (Clark) or an "identity statement" (Hardcastle), asserting that consciousness is identical to some physical or functional state. Now, it is certainly true that *if* we augment an account of physical processes with an identity statement of this form, the existence of consciousness can be derived; and with a sufficiently detailed and systematic identity statement, detailed facts about consciousness might be derived. But the question is now: what is the relationship between the physical facts and the identity statement itself?

Neither Clark nor Hardcastle gives us any reason to think that the identity statement *follows* from the physical facts. When answering the question "why does this physical process give rise to consciousness?", their answer is always "because consciousness and the physical process are identical", where the latter statement is something of a primitive. It is inferred to explain the correlation between physical processes and consciousness in the actual world, but no attempt is made to explain or derive it in turn. And without it, one does not come close to explaining the existence of consciousness.

This identity statement therefore has a very strange status indeed. It is a fact about the world that cannot be derived from the physical facts, and therefore has to be taken as axiomatic. No other "identity statement" above the level of fundamental physics have this status. The fact that DNA is a gene can be straightforwardly derived from the physical facts, as can the fact that H₂O is water, given only that one has a grasp of the concepts involved. Papineau (1996) argues that identities are not the sort of thing that one explains; I think this is wrong, but in any case they are certainly the kind of thing that one can *derive*. Even the fact that Samuel Clemens is Mark Twain, to use Papineau's example, could be derived in principle from the physical facts by one who possesses the relevant concepts. But even if one possesses the concept of consciousness, the identity involving consciousness is not derivable from the physical facts.

(It might be objected that if one possessed an *a posteriori* concept of consciousness - on which consciousness was identified with some neural process, for example - then the facts about consciousness could be derived straightforwardly. But this would be cheating: one would be building in the identity to derive the identity. In all other cases - genes, water, and so on - one can derive the high-level facts from the low-level facts using the *a priori* concept alone. One does not *need* the identity between genes and DNA to derive the fact that DNA is a gene, for example: all one needs is a grasp of the meaning of "gene". That is, in all the other cases, the implication from micro to macro is *a priori*.)

We might call this this "magic bullet" version of the identity theory: it treats identity as a magic bullet which one can use to kill off all our explanatory problems by drawing disparate phenomena together. But identities do not work like this: elsewhere, they have to be *earned*. That is, an identity requires an actual or possible explanation of how it is that two phenomena are identical. ("No identification without explanation.") One earns the DNA-gene identity, for example, by showing how DNA has all the properties that are required to qualify as a gene. The original identity theorists in the philosophy of mind (Place 1956; Smart 1959) understood this point well. They consequently buttressed their account with a "topic-neutral" analysis of experiential concepts, asserting that all it means to be an orange sensation is to be the sort of state caused by orange things, and so on; this suffers from all the problems of type-A materialism, but at least it recognizes what is required for their thesis to be true. The type-B materialist, by contrast, posits an identification *in place of* an explanation.

Indeed, type-B materialism seems to give up on the reductive *explanation* of consciousness altogether. The very fact that it needs to appeal to an explanatorily primitive axiom to bridge the gap shows that consciousness is not being wholly explained in terms of physical processes: a primitive bridging principle is carrying the central part of the burden, just as it does on the sort of theory I advocate. Calling this principle an "identity" may save the letter of materialism, but it does not save the spirit. When it comes to issues of *explanation*, this position is just as nonreductive as mine.

Elsewhere in science, this sort of explanatorily primitive link is found only in fundamental laws. In fact, this primitiveness is just what makes such laws fundamental. We explain complex data in terms of underlying principles, we explain those principles in terms of simpler principles, and when we can explain no further we declare a principle fundamental. The same should hold here: by positing a fundamental law, we recognize the price of explanatory primitiveness, rather than pretending that everything is business as usual.

One can draw out the problems in other ways. For example, once it is noted that there is no conceptually necessary link from physical facts to phenomenal facts, it is clear that the idea of a physically identical world *without* consciousness is internally consistent. (By comparison, a physically identical world without life, or without genes, or without water is not even remotely conceivable.) So the fact that physical processes go along with consciousness seems to be a *further* fact about our world. To use a common philosophical metaphor: God could have created our world without consciousness, so he had to do extra work to put consciousness in.

Type-B materialists sometimes try to get around this by appealing to Saul Kripke's treatment of *a*

posteriori necessity: such a world is said to be conceivable but not "metaphysically possible", precisely because consciousness is identical to a physical process. (Hardcastle embraces this line, and Clark says something similar). But as I argue in my book, this misunderstands the roots of *a posteriori* necessity: rather than ruling conceivable worlds impossible, *a posteriori* constraints simply cause worlds to be redescribed, and the problem returns as strongly as ever in a slightly different form. The issues are technical, but I think it is now well-established that Kripkean *a posteriori* necessity cannot save materialism here. To declare that the relevant worlds are all "metaphysically impossible", one would have to appeal instead to a far stronger notion of necessity which would put inexplicable constraints on the space of possible worlds. This is a notion in which we have no reason to believe.

So the problems of type-B materialism can be expressed both on intuitive and technical grounds. On the most intuitive grounds: it is a solution by stipulation, which "solves" the problem only by asserting that brain states are conscious states, without explaining how this can be. On slightly more technical grounds: it requires an appeal to a primitive axiom identifying consciousness with a physical process, where this identity is not derivable from the physical facts and is thus unlike any identity statement found elsewhere. On the most technical grounds: it either rests on an invalid appeal to Kripke's *a posteriori* necessity or requires a new and stronger notion of metaphysical necessity in which there is no reason to believe.

On to some specific points. **Clark** suggests that the explanatory gap arises only from assuming that consciousness and physical processes are distinct in the first place, and he faults my use of phrases such as "arises from" for begging that question. I think this misses the point: one can phrase the question just as well by asking "Why are certain physical systems conscious?", or even "Why is there something it is like to engage in certain processes?". Such questions are just as pressing, and clearly do not beg any questions against identity.

In fact, ontological assumptions are irrelevant to posing the explanatory question. All that matters is the *conceptual* distinction between structural/functional concepts and consciousness, a distinction that Clark explicitly accepts. (His talk of "correlation" makes even more clear, as does his observation that it *could* turn out that the functions do not correlate with experience). Given that it is not *a priori* that the performance of these functions should be conscious, it follows that an explanation of the functions is not ipso facto an explanation of consciousness, and we need to supplement the explanation with some further *a posteriori* component. Clark's "identity hypothesis" provides this extra component; but its primitive nature makes it clear that a wholly reductive explanation is not on offer. Indeed, Levine (1983), who introduced the term "explanatory gap", embraces an "identity" picture just like this, but he is under no illusion that he is providing a reductive explanation.

Hardcastle offers her own diagnosis of the roots of the debate, painting a picture of "committed materialists" who can't take the issue seriously, and "committed skeptics" who are entirely sure that materialism is false. I think this picture is far too bleak: in my experience the majority of people are more than a little torn over these issues, and there is plenty of common ground. In particular, I think Hardcastle does materialists a disservice: to characterize materialism as a "prior and fundamental commitment" is to make it into a religion. Materialism is an *a posteriori* doctrine, held by most because it explains so much in so many domains. But precisely because of this *a posteriori* character, its truth stands or falls with how

well it can explain the phenomena. So materialists cannot just circle the wagons and plead a prior commitment; they have to face up to the problems directly.

In any case, I think the basic intuitive divide in the field is not that between "materialists" and "skeptics", but that between those who think there is a phenomenon that needs explaining and those who think there is not: that is, between type-A materialists and the rest. The issue between Dennett and myself, for example, comes down to some basic intuitions about first-person phenomenology. But once one accepts that there is a phenomenon that needs explaining - as Hardcastle clearly does - the issues are more straightforwardly debatable. In particular, the problems of the type-B position are straightforwardly philosophical, rooted in its need for explanatorily primitive identities and brute metaphysical necessities.

Indeed, I think Hardcastle's defense of her identities makes straightforwardly philosophical missteps. Against someone who raises an explanatory gap question ("why couldn't these physical processes have gone on without consciousness?"), she responds with an analogy, pointing to a water-mysterian who asks "why couldn't water have been made of something else?", and a life-mysterian who asks "why couldn't living things be made from something other than DNA?". But such questions are disanalogous and irrelevant, as they get the direction of explanation backward. In reductive explanation, the direction is always from micro to macro, not vice versa. So even if life could have been made of something else, this blocks the DNA-explanation of life not in the slightest. What matters is that in these cases, the low-level facts imply the high-level facts, with no primitive identity statements required. But this is not so in the case of consciousness; so Hardcastle requires a primitive identity of an entirely different kind, for which analogies cannot help.

For a truly consistent type-B materialism, one would have to face up to these problems directly, rather than trying to slide over them. One would have to embrace explanatorily primitive identities that are logically independent of the physical facts and thus quite unlike any identities found elsewhere in science. One would have to embrace inexplicable metaphysical necessities that are far stronger than any *a posteriori* necessities found elsewhere in philosophy. And one will have to make a case that such postulates are a reasonable thing to believe in. I am skeptical about whether this is possible, but it is at least an interesting challenge.

But even if type-B materialism is accepted, the *explanatory* picture one ends up with looks far more like my naturalistic dualism than a standard materialism. One will have given up on trying to explain consciousness in terms of physical processes alone, and will instead be relying on primitive bridging principles. One will have to infer these bridging principles from systematic regularities between physical processes and phenomenological data, where the latter play an ineliminable role. One will presumably want to systematize and simplify these bridging principles as much as possible. (If there are to be brute identities in the metaphysics of the world, one hopes they are at least simple!) The only difference will be that these primitive principles will be called "identities" rather than "laws".

I think it makes far more sense to regard such primitive principles as laws, but if someone insists on using the term "identity", after a while I will stop arguing with them. In the search for a *theory* of consciousness - the truly interesting question - their theories will have the same shape as mine. The epistemology will be

the same, the methodology will be the same, the explanatory relations between principles and data will be the same, and all will be quite unlike those on standard materialist theories in other domains. The names may be different, but for all explanatory purposes, consciousness might as well be irreducible.

2.4 Other deflationary approaches

A different sort of "deflationary" approach is taken by **O'Hara and Scutt**. Their paper has the juicy title of "There is no hard problem of consciousness", suggesting a Dennett-like reductionism, but the substance of their paper suggests quite the opposite. In fact, they hold that the hard problem is so hard that we should ignore it for now, and work on the easy problems instead. Then perhaps all will become clear, in a decade or a century or two.

Now there is not much doubt that progress on the easy problems is much faster than progress on the hard problem, but O'Hara and Scutt's policy suggestion seems quite redundant. Researchers working on the easy problems already outnumber those working on the hard problem by at least a hundred to one, so there is not much danger of the world suddenly falling into unproductive navel-gazing. But if O'Hara and Scutt are suggesting that *no-one* should be working on the hard problem, this seems to move beyond pragmatism to defeatism. Granted that the hard problem is hard, it nevertheless seems quite reasonable for a community to invest a fraction of its resources into trying to solve it. After all, we do not *know* when a solution to the hard problem will come. Even if we do not solve it immediately, it may well be that the partial understanding that comes through searching for a solution will help us in the further search, in our work on the easy problems, and in our understanding of ourselves. It is in the scientific spirit to *try*.

Sociological issues aside, the substantive issue arising from O'Hara and Scutt's article is that of whether there is any chance of progress on the hard problem any time soon. O'Hara and Scutt do not really provide much argument against this possibility; they simply reiterate that the hard problem is very hard, that we are not assured of a solution, and that scientific progress has often made hard problems seem easier. All this tells us that the prospects for a solution are uncertain, but it does not tell us that they are nonexistent.

In my article I advocated a positive methodology for facing up to the hard problem. Pay careful attention both to physical processing and to phenomenology; find systematic regularities between the two; work down to the simpler principles which explain these regularities in turn; and ultimately explain the connection in terms of a simple set of fundamental laws. O'Hara and Scutt offer no reason to believe that this must fail. They reserve most of their criticism for reductive methods such as those of Crick and Edelman, but that criticism does not apply here. They very briefly criticize a specific suggestion of mine, saying "it is impossible to understand how information can have a phenomenal aspect". They do not substantiate this remark (for my part, I do not find it impossible to understand at all, as long as we realize that a fundamental law rather than a reduction is being invoked) but in any case the criticism seems quite specific to my theory. O'Hara and Scutt give us no reason to believe that a fundamental theory could not be formulated and understood.

I should also clarify a common misunderstanding. O'Hara and Scutt attribute to me the view that

understanding the easy problems does not help at all in understanding the hard problem, and others have attributed to me the view that neurobiology has nothing to contribute in addressing the hard problem. I did not make these claims, and do not agree with them. What I do say is that any account of the easy problems, and indeed any neurobiological or cognitive account, will be *incomplete*, so something more is needed for a solution to the hard problem. But this is not to say that they will play no role in a solution at all. I think it is obvious that empirical work has enriched our understanding of conscious experience a great deal, and I expect that it will continue to do so. A final theory of human consciousness will almost certainly lie in a combination of processing details and psychophysical principles: only using both together will the facts about experience be explained.

So I agree with O'Hara and Scutt that research on the easy problems is of the utmost importance: it is here that the meat and potatoes of consciousness research resides, and attention to this sort of work can help even a philosopher in staying grounded. But to ignore the hard problem entirely would be futile, as understanding conscious experience per se is the *raison d'être* of the field. Some of us will continue to focus on it directly, and even those working on the easy problems will do well to keep the hard problem in sight out of the corner of their eyes. To paraphrase Kant and stretch things a bit, we might say: hard without easy is empty; easy without hard is blind.

Another proposal that could be construed as "deflationary" comes from **Price**, who suggests that much of the problem lies in our heads. We should not expect to *feel* as if we understand consciousness, but this may be no big deal. There are similar explanatory gaps accompanying every causal nexus ("why does event A cause event B?"); it's just that in most cases we have gotten used to them. The explanatory gap in the case of consciousness is analogous, but we are not yet as used to it.

I agree with Price's analogy, but I think it ultimately supports my view of the problem. Why are causal nexi accompanied by explanatory gaps? Precisely because of their contingency (as Price says, there is no "*a priori* necessity" to them), which is in turn due to the brute contingency of fundamental laws. If we ask "why did pressing the remote control cause the TV set to turn on?", we might get a partial answer by appealing to principles of electromagnetic transmission, along with the circuitry of the two objects, ultimately seeing how this causal chain is the natural product of the underlying dynamics of electromagnetism (for example) as it applies to the material in the vicinity. But this answer is only partial, as we have no answer to the question of "why do those fundamental principles hold?". Those principles are apparently just a brutally contingent fact about the world, and this contingency is inherited by the causal chain at the macroscopic level.

If Price is right that the explanatory gap between brain and consciousness is analogous, then this suggests that the gap is due to some contingency in the connecting principles, because of underlying brutally contingent fundamental laws. Which of course is just what I suggest. We have here an inter-level relationship that *could have been otherwise*, just as Price points to intra-level relationships in physics that could have been otherwise. Either way, this arbitrariness is ultimately grounded at the point where explanation stops: the invocation of fundamental laws.

It is worth noting that for other inter-level relationships - that between biochemistry and life, for example,

or between statistical mechanics and thermodynamics - there is no explanatory gap analogous to the brain-consciousness gap. The reason is precisely that the high-level facts in these cases are *necessitated* by the low-level facts. The low-level facts themselves may be contingent, but there is no further contingency in the inter-level bridge. (Indeed, the inter-level relationship in these cases is not really causation but *constitution*.) Because there is no contingency here, the relationship between the levels is transparent to our understanding. Contrapositively, the lack of transparency in the brain-consciousness case is precisely due to the contingency of the psychophysical bridge.

In any case, Price's analogy between the brain-consciousness relation and ordinary causal relations is helpful in seeing why belief in an explanatory gap need not lead one to mysterianism. Rather than elevating the explanatory gap to a *sui generis* mystery, we recognize that it is of the sort that is ubiquitous elsewhere in science, and especially in fundamental physics. This case is unusual only in that here, the gap is found in an inter-level rather than an intra-level relationship; but the same strategy that works for intra-level relationships works here. Once we introduce fundamental psychophysical laws into our picture of nature, the explanatory gap has itself been explained: it is only to be expected, given that nature is the way it is.

A final view that might be considered "deflationary" has been discussed by **McGinn**, not so much in his contribution to this symposium as in an earlier paper (McGinn 1989) and most explicitly in his review of my book (McGinn 1996). On McGinn's view, the explanatory gap also arises for psychological reasons, but his reasons differ from Price's. He suggests that there may be a conceptual implication from physical facts to facts about consciousness, which would be *a priori* for a being that possessed the relevant concepts; but we do not and cannot possess the concepts, due to our cognitive limitations, so we can never grasp such an implication. On this view, materialism turns out to be true, but we can never grasp the theory that reveals its truth.

This intriguing view seems at first glance to offer an attractive alternative to both dualism and hard-line reductionism, but in the end, I am not sure how much of an alternative is. The problem lies in the concept (or concepts) which support the implication from physical to phenomenal facts. What sort of concept could this be? If it is a structural/functional concept, then it will suffer from the same conceptual gap with experiential concepts as any other structural/functional concept (the existence of a gap here is independent of *specific* details about structure and function, after all). If it is not a structural/functional concept, then there appear to be principled reasons why it cannot be entailed by the physical story about the world, as physics deals only in structure and function.

So we are still faced with the problem that structure and function adds up only to more structure and function. This claim holds true for systematic reasons quite independent of considerations about cognitive limitations, and I doubt that McGinn would deny it. So it seems that McGinn need to assert either that (1) explaining experience *is* just a problem of explaining structure and function, if only we could grasp this fact, or (2) that something more than structure and function is present in fundamental physics. The first option would make McGinn's position remarkably like **Dennett's** (the only difference being that Dennett holds that only *some* of us are limited in this way!), and the second position would fall into the category of expanding fundamental physics, which I will consider below. Either way, once made specific, this view

is subject to the pros and cons of the specific position to which it is assimilated. So in the end, it may not open up a distinct metaphysical option.

3 NONREDUCTIVE ANALYSES

3.1 Conceptual foundations

I will now address some critiques from those who take nonreductive positions. It appears that I staked out some middle ground; having discussed objections from my right, it is now time for objections from the left. The intermediate nature of my position may stem from an inclination toward simplicity and toward science. Reductive materialism yields a compellingly simple view of the world in many ways, and even if it does not work in the case of consciousness, I have at least tried to preserve as many of its benefits as possible. So where reductionists think that I have overestimated the difficulty of the hard problem, some nonreductionists think that I may have underestimated it, or alternatively that I have underestimated the difficulty of the "easy" problems.

The latter position - that the hard problem is hard, but that explaining discrimination, reportability, and so on is just as hard - is taken by **Lowe** and **Hodgson**, for two apparently different reasons. Hodgson thinks these problems are hard because a physical account cannot even explain how the functions are performed; Lowe thinks they are hard because they require explaining more than the performance of functions. (It is possible that Lowe intends to make both points.) I will address Lowe's position first, and save Hodgson's for my discussion of interactionism and epiphenomenalism.

Why say that explaining reportability, discrimination, and so on requires explaining more than the performance of functions? Lowe says this because he holds that true "reports" and "discriminations" can be made only in systems which have the capacity for thought, which in turn requires consciousness. If externally indistinguishable functions were performed in a system without consciousness, they would qualify as "reports" (and so on) only in a "jejune" sense. So an account of non-jejeune reportability requires explaining more than functions.

I have some sympathy with Lowe's position here; in particular, I find it plausible that there is an intimate relationship between consciousness and thought (Lowe suggests that I think otherwise, but I don't think that suggestion can be found in my article). But it seems to me that the issue about "reportability" and so on is largely verbal. Does a sound uttered by a functionally identical zombie really qualify as a "report"? The answer is surely: yes in one sense of "report", and no in another. If Lowe objects to calling it a "report" in any sense at all, one can simply call it a "pseudo-report". Then the easy problems are those of explaining pseudo-reportability, pseudo-discrimination, and the like. Nothing important to my article changes; the distinction between the easy problems (explaining functions) and the hard problem (explaining conscious mentality) is as strong as ever.

Lowe might reply that reportability, so construed, is not a problem of *consciousness* at all. Again I am sympathetic, but again I think that this is a verbal issue. Plenty of people who take functional approaches

to these problems take themselves to be explaining aspects of consciousness in some sense; and there is little point getting into territorial arguments about a word. It's more productive to accept the characterization - if someone holds that "consciousness" has *some* functionally definable senses, I will not argue with them - but to point out the key problems of consciousness that are being skipped over all the same.

The same goes for Lowe's concerns, shared by **Velmans** and **Libet**, about my use of the term "awareness" for a functionally defined concept distinct from that of full-blown consciousness. Again, a word is just a word. As long as we are clear that "awareness" is being used in a stipulative sense, the substantive issues should be clear. In particular, there is certainly no implication that humans are "aware" *only* in this attenuated sense, as Lowe somehow infers; and it is hard to see how this terminological choice helps blur the "function/sentience" distinction, as Velmans suggests. If anything, explicitly separating consciousness and awareness makes the distinction harder to avoid. Nevertheless, it is clear that enough people are uneasy about the terminology that it is unlikely to catch on universally. Perhaps another term can play the role, although I suspect that any word choice that dimly suggests mentality would meet similar opposition from some. It's a pity that there is no universal term for this central functional concept; in the meantime I will go on using the term "awareness", with the stipulative nature of the usage always made clear.

The exact relationship between consciousness and "intentional" (or semantic) mental states such as belief, thought, and understanding raises deep and subtle questions that I did not intend to address in my article. Lowe seems to have gotten the impression of a straightforward functionalism about these aspects of mentality, but such an impression was not intended. I am torn on the question of intentionality, being impressed on one hand by its phenomenological aspects, and on the other hand being struck by the potential for functional analyses of specific intentional contents. In my book, I try to hew a neutral line on these deep questions, noting that there is a "deflationary" construal of concepts such as "belief" so that even a zombie might be said to have beliefs (pseudo-beliefs, if you prefer), and an inflationary construal such that true belief requires consciousness. Over time I am becoming more sympathetic with the second version: I think there may be something in the intuition that consciousness is the primary source of meaning, so that intentional content may be grounded in phenomenal content, as Lowe puts it. But I think the matter is far from cut and dried, and deserves a lengthy treatment in its own right. For now, phenomenal content is my primary concern.

3.2 The roots of the hard problem

Robinson, **McGinn**, and **Warner** offer proposals about why the hard problem is hard. These are not direct critiques of my view, for the most part, but they fall into the general category of nonreductive analyses, so I will address them briefly here.

Robinson suggests that the hardness lies in the fact that some phenomenal properties - hue properties, for example - have no structural expression. I think there is a considerable insight here. Elsewhere in science, instantiations of structural properties are generally explicable in terms of basic components and their relations, and it seems to be precisely their structure that makes them explicable in this way. The

structural properties of experience itself (the geometry of a visual field, for example) form an interesting intermediate case: while they are more amenable to physical explanation than other phenomenal properties, this explanation still requires a nonreductive principle to cross the gap. But these properties may be reducible to structureless *phenomenal* properties and their relations. If so Robinson may be correct that the core of phenomenal irreducibility lies at the more basic level.

A few questions remain: for example, if it turned out that phenomenal properties had structure "all the way down", might they not be irreducible to physical properties all the same? For reasons like this, I sometimes lean toward an alternative view which locates the irreducibility in an independence of the *kind* of structure found in the physical domain, and ultimately in the *intrinsicness* of phenomenal properties, which contrasts with the relational nature of all our physical concepts. But clearly these views are not far apart.

McGinn offers a closely related analysis. He locates the problem in the *non-spatial* character of consciousness: that is, in the fact that it lacks spatial extension and structure, and therefore does not fit easily into physical space. I think the overall intuition is very powerful. The detailed claim needs to be carefully unpacked, to avoid lumping in consciousness with less problematic non-spatial states and properties (e.g. the legality of an action, which is a complex dispositional property but not a spatial property; and possibly even the charge of a particle), while simultaneously avoiding the need to appeal more controversially to a non-spatial *entity* bearing the state or property (McGinn seems not to want to rest his case on this appeal; see e.g. his footnote 3). I suspect that once this work is done - adding appropriate restrictions on the class of properties, perhaps - McGinn's analysis will be even closer to those above.

Warner locates the source of the problem in a different place: the *incorrigibility* of our knowledge of consciousness. I agree with Warner that there is some sense in which some knowledge of consciousness is incorrigible - I know with certainty that I am conscious right now, for example - but it is remarkably tricky to isolate the relevant sense and the relevant items of knowledge. Warner himself notes that plenty of our beliefs about our experiences are mistaken. He gets around this problem by limiting this to cases where our ability to recognize experiences is "unimpaired", but this seems to come dangerously close to trivializing the incorrigibility claim. After all, it is arguably a tautology that an "unimpaired" belief about an experience will be correct. Warner may have a way to unpack the definition of "impairment" so that the claim is non-circular, but this is clearly a non-trivial project.

In Chapter 5 of my book (pp. 207-8), I make some brief suggestions about how to make sense of an incorrigibility claim. In essence, I think that experiences play a role in constituting some of our concepts of experience; and when a belief directs such a concept at the experience which constitutes it, there is no way that the belief can be wrong (in essence, because one's current experience has got "inside" the content of one's belief). Many or most beliefs about experience do not have this specific form, and are therefore corrigible; nevertheless, this may isolate a certain limited class of beliefs about experience that cannot be wrong. (This limited class of beliefs can arguably ground the first-person epistemology of conscious experience, but this is a further complex issue.)

In any case, Warner and I are agreed that there are *some* beliefs about conscious experience that cannot be wrong. What follows? Warner holds that it follows from this alone that experience cannot be physically explained, as physical science cannot countenance the necessary connections that incorrigibility requires. I am not sure about this. On my account, for example, the necessary connection between belief and experience is an automatic product of the role that the experience plays in constituting the content of the belief; and it is not *obvious* to me that materialists could not avail themselves of a similar account. Shoemaker (1990) gives an alternative account of incorrigibility from a functionalist perspective, relying on the interdefinition of pains and pain-beliefs. Perhaps Warner would object that neither of these accounts captures the kind of incorrigibility that he is after; but perhaps they capture the kind of incorrigibility in which there is reason to believe. So I am not yet convinced that incorrigibility is truly the source of the mind-body problem, but it is clear that there is much more to be said.

Warner uses these considerations about incorrigibility to suggest, like Lowe, that even reportability - one of my "easy" problems - cannot be physically explained. My reply here is as before. I did not intend reportability to be read in a strong sense that *requires* the presence of experience. Rather, I intended it to require merely the presence of the reports, functionally construed, so in particular I did not intend it to encompass the incorrigibility of beliefs about experience. (If I were writing the article now, I would modify the wording in the list of "easy" problems to make it absolutely clear that functioning is all that matters.) Certainly, *if* "reportability" is read in a sense that requires conscious experience, then it cannot be reductively explained.

3.3 Fundamental laws

A further set of issues is raised by my appeal to fundamental laws in a theory of consciousness. **Mills** thinks that because I invoke such laws to bridge physics and consciousness, I am not really solving the hard problem at all (**Price** suggests something similar). At best I am providing a sophisticated set of correlations, and finding such correlations was an easy problem all along.

Mills reaches this conclusion because he construes the hard problem as the problem of giving a *constitutive* (or "non-causal") explanation of consciousness in physical terms. If the problem is construed that way, Mills is quite right that it is not being solved at all. But to define the problem of consciousness this way would be to define it so that it becomes unsolvable: one might call *that* problem the "impossible problem".

I prefer to set up the hard problem in such a way that a solution is not defined out of existence. The hard problem, as I understand it, is that of explaining how and why consciousness arises from physical processes in the brain. And I would argue the sort of theory I advocate can in principle offer a good solution to this problem. It will not solve the impossible problem of providing a reductive explanation of consciousness, but it will nevertheless provide a theory of consciousness that goes beyond correlation to explanation.

A good analogy is Newton's theory of gravitation. The Newton of legend wanted to explain why an apple

fell to the ground. If he had aimed only at correlation, he would have produced a taxonomic theory that noted that when apples were dropped from such-and-such heights, they fell to the ground taking such-and-such time, and so on. But instead he aimed for *explanation*, ultimately explaining the macroscopic regularities in terms of a simple and fundamental gravitational force. In Newton's time, some objected that he had not explained why the gravitational force should exist; and indeed he had not. But we take Newton's account to be a good explanation of the apple's falling all the same. We have grown used to taking some things as fundamental.

Something similar holds for a theory of consciousness. It would be deeply unsatisfying for a theory of consciousness to stop at "complex brain state B is associated with complex experience C", and so on for a huge array of data points. As in Newton's case, we want to know how and why these correlations hold; and we answer this question by pointing to simple and fundamental underlying laws. Just as one can say "the apple fell because of the law of gravity", we will eventually be able to say "brain state B produced conscious state C because of fundamental law X".

Because something is being taken as primitive, this does not yield as *strong* an explanatory connection as one finds in cases of reductive explanation, such as the explanation of genes in terms of DNA. But it is an explanation all the same. The case of gravity suggests that what counts in an explanation is that one reduces the primitive component to something as simple as possible, not that one reduces it to zero.

Mills suggests that this is no better than explaining why a sheep is black in terms of the fact that it is a member of the class of black things. But here the explanatory posit is just as complex as what needs to be explained; whereas in our case, the fundamental laws are far simpler than the data. If our "explanation" was "brain B yields experience E" or even "certain oscillations yield consciousness", we would have a problem like Mills': these posits would be so complex and macroscopic that they stand in need of further explanation themselves. For a comprehensive explanation, our basic principles need to be so simple and universal that they are plausibly part of the basic furniture of the world.

Of course one can always ask "why does the fundamental law hold", as Mills and also **Robinson** suggest. But we should not expect any answer to that question. In physics, we have grown used to the idea that explanation stops somewhere, and that the fundamental laws of nature are not further explained. That is what makes them fundamental. If my negative arguments about consciousness are correct, then we will have to do the same here. We will explain and explain and explain, and eventually our psychophysical explanations will be reduced to a simple core which we will take as primitive. So we do not get something for nothing, but we get a perfectly adequate theory all the same.

Mills is right that once we view things this way, there is a sense in which the hard problem becomes an easy problem (although not an Easy problem), in that there is a clear research program for its solution and there is no reason why it should be intractable in principle. This I take to be precisely the liberating force of taking consciousness as fundamental. We no longer need to bash our head against the wall trying to reduce consciousness to something it is not; instead we can engage in the search for a constructive explanatory theory.

In any case, it seems that Mills does not disagree with me on the issues of substance. Whichever problems one takes to be "hard" or "easy", the deepest problem of consciousness is that of how we can construct an explanatory theory of consciousness which accommodates consciousness in the natural world. And a fundamental theory of consciousness, we agree, is the best way to do just that. I will be happy if we can come up with a theory of consciousness that is only as good as Newton's theory of gravitation!

3.4 Epiphenomenalism and interactionism

A number of contributors worry that my position may lead to epiphenomenalism, the view that consciousness has no effect on the physical world. If the physical domain is causally closed, so that there is a physical explanation for every physical event, and if consciousness is non-physical, then it can seem that there is no room for consciousness to play any causal role. Conversely, it can seem that if consciousness is non-physical and plays a causal role, then there will not be a physical solution even to the "easy" problems. **Hodgson** and **Warner** spend some time discussing this issue, and **Seager** and **Stapp** allude to it. I discuss this issue at considerable length in my book, but will summarize the state of play as I see it below.

In essence, I think that (1) while epiphenomenalism has no clear fatal flaws, it is to be avoided if possible; that (2) the causal closure of the physical domain is not to be denied lightly; and that (3) denying causal closure does not really help solve the problems of epiphenomenalism, which run deeper than this. Most importantly, I think that (4) it may be possible to avoid epiphenomenalism even while embracing the causal closure of the physical domain, by taking the right view of the place of consciousness in the natural order. I will consider these issues in order.

First, is epiphenomenalism an acceptable view, or should it be rejected out of hand? There is no doubt that the view is counterintuitive to many, but it is also hard to find fatal flaws in it. While we certainly have strong intuitions that consciousness plays a causal role, our *evidence* for these intuitions lies largely in the fact that certain conscious events tend to be systematically followed by certain physical events. As always, when faced with such a constant conjunction, we infer a causal connection. But the epiphenomenalist can account for this evidence in a different way, by pointing to psychophysical laws, so our intuitions may not carry too much weight here.

Hodgson argues vigorously against epiphenomenalism, largely by appealing to "common sense". I think common sense should not be undervalued here, but it is also inconclusive. At best, it establishes a presumption against epiphenomenalism if other things are equal, not a solid argument against it if other things are not. Hodgson also points to various functions that he thinks could not be performed as well without consciousness; but his arguments all depend once again on the intuition that consciousness is playing a causal role, rather than on an objective analysis of the functions themselves. He also makes an appeal to evolution, but an epiphenomenalist can account for the evolution of consciousness without too many problems: evolution selects for certain physical processes directly, and psychophysical laws do the rest, ensuring that consciousness will evolve alongside those processes. Like all fundamental laws, these psychophysical laws are universal, so we do not need an evolutionary explanation of why these laws hold

in the first place.

Other anti-epiphenomenalist arguments can be made by appealing to the relationship between consciousness and the things we *say* and *judge* about consciousness. It seems that the epiphenomenalist must hold that consciousness is causally irrelevant to our utterances about consciousness, which is at least very odd. Some argue that it is more than odd, suggesting that if consciousness were epiphenomenal we could not *refer* to consciousness, or that we could not *know* about consciousness; but I think that a close analysis, as I give in my book, suggests that these arguments do not go through, as our knowledge of and reference to consciousness depends on a relationship to consciousness that is much tighter than mere causation.

Warner gives a novel argument against epiphenomenalism, and against any other view that has a causally closed physical domain plus psychophysical laws. He suggests that psychophysical laws *must* interfere with physical laws, as they automatically entail violations of physical conservation laws. I do not see why this is the case: surely it is at least coherent to suppose that the physical picture of the universe might be supplemented by some psychophysical laws that introduce consciousness but leave the physical domain untouched, Warner's argument relies on the claim that the "production" of experience by a physical process must involve a corresponding decrease in some physical quantity, but I see no reason why this must be so: there will be *some* physical criterion for the existence of an experience, to be sure, but this criterion may be one that can be satisfied perfectly well in a causally closed physical world. So the conceptual coherence of epiphenomenalism, and that of other views with causal closure plus psychophysical laws, is unthreatened.

Still, all this establishes at best that epiphenomenalism has no fatal flaws. It does not establish that epiphenomenalism is plausible. Not only does epiphenomenalism violate certain aspects of common sense; it also leads to an inelegant picture of nature, with consciousness "dangling" on top of physical processes as a kind of add-on extra. If it turns out that every other position has fatal flaws, then we may have reason to embrace epiphenomenalism; but in the meantime, we have good reason to investigate alternatives.

There are two sorts of alternatives that one might consider. First, we might see if it is plausible to deny the causal closure of the physical domain, thus leaving room for a causal role for experience in an interactionist dualism. Second, we might see if a causal role for experience might be *reconciled* with the causal closure of the physical domain. The second alternative may sound paradoxical at first, but I think there is a very natural way to make sense of it, which may ultimately provide the deepest resolution of this issue.

But first: is the physical world causally closed? In the paper I accepted that it was, not because I think things *have* to be that way, but because to deny this is to go a long way out on a limb. One does not have to go out on that limb to embrace the irreducibility of consciousness, so I prefer to stay neutral, lest the baby of consciousness once more be thrown out with the bathwater of Cartesian dualism. Still, are there any good reasons to deny causal closure, and to assert that physical explanations of the various functions are incomplete?

Perhaps the most common such reason is an indirect one: "it *must* be the case that physical explanations of the functions are incomplete, if consciousness is to play a causal role." This reason has some force, although I think both of its premises can be questioned: we have seen above that it is not obvious that consciousness *must* have a causal role, and we will see below that consciousness might have a causal role even if the physical domain is causally closed. But in any case, I set this indirect reason aside: the question for now is whether there are any *direct* reasons. That is, if we set consciousness aside and take a third-person view of the world, is there any reason to believe that physical explanations of these functions are impossible?

Hodgson offers an array of reasons to deny causal closure, but they are mostly grounded in the indirect reason above. Hodgson does not deny that some physical system might perform the functions with which the "easy" problems are concerned; he simply thinks that that is not the way that *we* do it, as consciousness plays a role in our own case. So his case that the "easy" problems are hard depends largely on the existence of the hard problem, and not on considerations intrinsic to the easy problems themselves. Indeed, I think that "objective" reasons suggesting that no physical systems could perform these functions are very thin on the ground.

The main place where third-person considerations may give reason to deny causal closure is in the intriguing case of quantum mechanics, which both Hodgson and **Stapp** appeal to. While there are interpretations of quantum mechanics on which the physical domain is causally closed - the interpretations of Bohm and Everett, for example - there are also interpretations on which it is not, and which leave a potential causal role for consciousness wide open. Stapp, for example, favors an interpretation on which consciousness is responsible for "collapsing" the wave function, and Hodgson favors an interpretation on which consciousness determines certain apparent quantum indeterminacies.

Indeed, it can seem that quantum mechanics provides about as perfect a causal role for consciousness as one could imagine in a physical theory. Any indeterminism in quantum mechanics comes in at the point of "collapse", which on the most common interpretations is triggered by "measurement", and it can seem that consciousness is the only non-arbitrary way to distinguish a measurement from other physical events. If so, then consciousness may be present in quantum mechanics' very foundations. Such interpretations are controversial among physicists, but mainly because they presuppose that consciousness is non-physical; if we have already accepted this for independent reasons, this concern loses its bite. (It is interesting that philosophers reject interactionist dualism because they think it is incompatible with physics, whereas physicists reject the relevant interpretations of quantum mechanics because they are dualistic!)

On most days of the week, I lean toward a different interpretation of quantum mechanics (Everett's), but interactionist collapse interpretations have obvious attractions and are not to be dismissed lightly. (I lean toward them about two days a week, and toward Bohm's interpretation on Sundays.) At least it seems clear that interactionist dualism is not *incompatible* with physical theory, as we understand it today. But I think there is a deeper reason why an appeal to interactionist dualism does not really solve the problems of epiphenomenalism. This is because even interactionism is subject to an epiphenomenalist worry of its

own! Perhaps it can get around this worry, but it turns out that the same move is available to theories on which physics is causally closed.

The worry is as follows: for any given interactionist theory, it seems that we can remove the facts about experience, and still be left with a coherent causal story. Take Eccles' theory on which "psychons" in the mind affect physical processes in the brain. Here one can tell a perfectly coherent causal story about psychons and their effect on the brain without ever mentioning the fact that psychons are *experiential*. On this story, psychons will be viewed as causal entities analogous to electrons and protons in physical theories, affected by certain physical entities and affecting them in turn; and just as with protons and electrons, the fact that psychons have any experiential qualities will be quite inessential to the dynamic story. So one can still give a causal explanation of behavior that does not involve or imply experience. The same would go for a Cartesian theory involving ectoplasm, for **Libet's** proposal involving a "conscious mental field", and even for the theories that Stapp and Hodgson advocate.

Consider **Stapp's** view, for example. Presumably when this view is filled out, it will say that certain physical states P give rise to certain experiential states E, and that these states E bring about physical collapses in turn. But however this story works, the fact that the states E are *experiential* will be quite inessential to the story. One can imagine that a formally identical theory might be formulated from a "God's-eye" point of view, invoking such states E in causing collapses, but never mentioning experience at all. So it is not easy to see how Stapp is giving experience an *essential* role.

Stapp has sometimes advocated his view by pointing to the "zombie" possibility for classical physics: if physics is causally closed, there is a logical possibility of physically identical zombies with the same behavior, suggesting that experience plays no essential role in our behavior. But interestingly, a similar objection can be made to Stapp's own view. Given that physics works as Stapp suggests, there is a logically possible world with a "quantum zombie". In this world, instead of P causing experience E which causes collapse, P causes collapse directly. There is no consciousness in this world, but all the functions are performed just the same. So there is a sense in which the fact that *experience* is associated with collapses in our world is superfluous. One can tell a similar conceptually coherent "zombie" story for any interactionist picture, whether Hodgson's or Eccles - just move to a possible world in which any intermediate causal roles are played without any associated experience -- thus suggesting that these problems are not unique to the picture on which the physical world is causally closed.

The real "epiphenomenalism" problem, I think, does not arise from the causal closure of the physical world. Rather, it arises from the causal closure of the world! Even on an interactionist picture, there will be *some* broader causally closed story that explains behavior, and such a story can always be told in a way that neither includes nor implies experience. Even on the interactionist picture, we can view minds as just further nodes in the causal network, like the physical nodes, and the fact that these nodes are experiential is inessential to the causal dynamics. The basic worry arises not because experience is logically independent of physics, but because it is logically independent of causal dynamics more generally.

The interactionist has a reasonable solution to this problem, I think. Presumably, the interactionist will respond that some nodes in the causal network are experiential through and through. Even though one *can*

tell the causal story about psychons without mentioning experience, for example, psychons are *intrinsically* experiential all the same. Subtract experience, and there is nothing left of the psychon but an empty place-marker in a causal network, which is arguably to say there is nothing left at all. To have real causation, one needs something to do the causing; and here, what is doing the causing is experience.

I think this solution is perfectly reasonable; but once the problem is pointed out this way, it becomes clear that the same solution will work in a causally closed physical world. Just as the interactionist postulates that some nodes in the causal network are intrinsically experiential, the "epiphenomenalist" can do the same.

Here we can exploit an idea that was set out by Bertrand Russell (1926), and which has been developed in recent years by Grover Maxwell (1978) and Michael Lockwood (1989). This is the idea that *physics* characterizes its basic entities only extrinsically, in terms of their causes and effects, and leaves their intrinsic nature unspecified. For everything that physics tells us about a particle, for example, it might as well just be a bundle of causal dispositions; we know nothing of the entity that *carries* those dispositions. The same goes for fundamental properties, such as mass and charge: ultimately, these are complex dispositional properties (to have mass is to resist acceleration in a certain way, and so on). But whenever one has a causal disposition, one can ask about the categorical *basis* of that disposition: that is, what is the entity that is doing the causing?

One might try to resist this question by saying that the world contains *only* dispositions. But this leads to a very odd view of the world indeed, with a vast amount of causation and no entities for all this causation to relate! It seems to make the fundamental properties and particles into empty placeholders, in the same way as the psychon above, and thus seems to free the world of any substance at all. It is easy to overlook this problem in the way we think about physics from day to day, given all the rich details of the mathematical structure that physical theory provides; but as Stephen Hawking (1988) has noted, physical theory says nothing about what puts the "fire" into the equations and grounds the reality that these structures describe. The idea of a world of "pure structure" or of "pure causation" has a certain attraction, but it is not at all clear that it is coherent.

So we have two questions: (1) what are the intrinsic properties underlying physical reality?; and (2) where do the intrinsic properties of experience fit into the natural order? Russell's insight, developed by Maxwell and Lockwood, is that these two questions fit with each other remarkably well. Perhaps the intrinsic properties underlying physical dispositions are themselves *experiential* properties, or perhaps they are some sort of *proto-experiential* properties that together constitute conscious experience. This way, we locate experience *inside* the causal network that physics describes, rather than outside it as a dangler; and we locate it in a role that one might argue urgently needed to be filled. And importantly, we do this without violating the causal closure of the physical. The causal network itself has the same shape as ever; we have just colored in its nodes.

This idea smacks of the grandest metaphysics, of course, and I do not know that it has to be true. But if the idea *is* true, it lets us hold on to irreducibility and causal closure and nevertheless deny epiphenomenalism. By placing experience inside the causal network, it now carries a causal role. Indeed,

fundamental experiences or proto-experiences will be the *basis* of causation at the lowest levels, and high-level experiences such as ours will presumably inherit causal relevance from the (proto)-experiences from which they are constituted. So we will have a much more integrated picture of the place of consciousness in the natural order. [*]

*[[[There may be other ways to reconcile a causal role for experience with the causal closure of the physical. See Mills (1995) for a different strategy that relies on causal overdetermination. But even if this view avoids epiphenomenalism, it retains a fragmented, inelegant picture of nature.]]]

The Russellian view still qualifies as a sort of "naturalistic dualism", as it requires us to introduce experience or proto-experience as fundamental, and it requires a deep duality between the intrinsic and extrinsic features of physical reality. But underlying this dualism, there is a deeper monism: we have an integrated world of intrinsic properties connected by causal relations. The view can even be seen as an odd sort of "materialism", as it says that physical reality is all there is - but it says that there is much more in physical reality than physical theory tells us about! In the end the name does not matter too much, as long as the picture is clear. (I would be tempted by "fundamentalism" as the most accurate coverall for the sorts of view I embrace, were it not for the associations!)

There are obvious concerns about this view. The first is the threat of panpsychism, on which more later. The second is the problem of how fundamental experiential or proto-experiential properties at the microscopic level somehow together *constitute* the sort of complex, unified experience that we possess. (This is a version of what **Seager** calls the "combination problem".) Such constitution is almost certainly required if our own experiences are not to be epiphenomenal, but it is not at all obvious how it should work: would not these tiny experiences instead add up to a jagged mess? I discuss some approaches to this problem later. *If* it can be avoided, then I think the Russellian view (which turns out to be particularly compatible with an informational "it from bit" view) is clearly the single most attractive way to make sense of the place of experience in the natural order.

It is notable that even an interactionist dualism can be seen as a sort of Russellian view. It draws a slightly different picture of the causal network, and takes certain nodes in this network - the "psychon" or "collapse" nodes, for example - and colors them in. The differences are that not *all* nodes in the network are colored in in this way (presumably there are some different, unknown intrinsic properties in fundamental matter), and that the experiential nodes in this picture are at a fairly high level. This may actually help avoid the problem above: instead of trying to constitute our consciousness out of innumerable *different* fundamental nodes, there might turn out to be a *single* node in each case (or just a few?) which carries the burden. (Though one may well wonder why this single node should have such a complex of intrinsic properties, in the way that our consciousness does!) This avoidance of the constitution problem may in the end turn out to be the greatest virtue of a quantum interactionism.

In the meantime, I think this question is wide open. There are at least three potential ways of seeing the metaphysics here: the epiphenomenalist version, the interactionist version, and the Russellian version. All have pros and cons, and I think the question of their mutual merits is one that deserves much further investigation.

3.5 My psychophysical laws

A few contributors made comments on the three specific proposals I made about psychophysical laws: the principle of structural coherence, the principle of organizational invariance, and the double-aspect view of information. Taking these in turn:

(1) The principle of structural coherence. This is the least controversial of the three proposals, and unsurprisingly there was not much argument with it. It has long been recognized that there is a detailed correspondence between structural properties of the information processed in the brain and structural properties of conscious experience (see the "psychophysical axioms" of Muller 1896 and the "structural isomorphism" of Kohler 1947, for example). My slightly more specific proposal, specifying that the relevant information is that made available for global control, is also implicitly or explicitly present in much current research.

The only criticism is by **Libet**, who thinks that my equation of the structure of consciousness with the structure of awareness is either trivial or false. I think he is placing too much weight on the use of the word "awareness" here, however; I use the term stipulatively to refer to global availability of information (availability for such processes as verbal report, among other things), and might easily have used another term instead. I suspect that when this verbal issue is set aside, Libet will not find much to disagree with.

(2) The somewhat functionalist principle of organizational invariance, and my arguments for it, met with a bit more disagreement. **Velmans** objects to it on the grounds that a cortical implant might produce a refined version of blindsight, with excellent performance but no verbal reports of consciousness and hence no experience. But this is no counterexample to the principle: the very absence of verbal reports in these subjects show that they are functionally inequivalent to normal subjects. Perhaps they are "functionally equivalent" in some very loose sense, but the invariance principle requires a much stricter isomorphism than this. The moral is that the processes involved in the production of verbal reports are just as much part of a subject's functional organization as the processes responsible for discrimination and motor action. Indeed, these aspects of organization may be among the prime determinants of conscious experience.

Similarly, **Libet** says that I rely on a "behavioral" criterion for conscious experience, instead of more convincing criteria such as a subject's verbal report. But a verbal report is a sort of behavioral criterion in its own right; and in any case, it is clear that any subject who is functionally isomorphic to me in the strict sense that the principle requires will produce exactly the same verbal reports, and so will satisfy Libet's criterion. Libet is quite right that there are cases where performance on many tasks is dissociated from verbal report, but such cases are irrelevant to assessing the principle.

A fairly common reaction to these thought-experiments is to suggest that no silicon chip could in fact duplicate the function of a neuron, or at least that one should not beg that question. I agree that this is clearly an open empirical question. The principle says only that *if* a system is a functional isomorph of a conscious system, it will have the same sort of experiences; it makes no claims about just how such isomorphs might be realized. Silicon chips are just an example. If silicon isomorphs turn out to be possible, then the principle applies to them; if they do not, the scope of the principle will be more limited. Either way, the idea that functional organization fully determines conscious experience is unthreatened by this line of questioning.[*]

*[[[That being said: if the laws of physics are computable, a neuron's behavior is in principle computable too, and it is not implausible that the relevant computations could be hooked up to electrical and chemical mediators with other neurons, at least in principle if not easily in practice. We already have seen artificial hearts, and people are working on artificial retinas; my own money is on the eventual possibility of artificial neurons.]]]

Hardcastle wonders if we can *really* know what will happen upon duplicating neural function in silicon. Here the answer is no and yes. No, we can't know for sure that neural function can be duplicated perfectly in silicon - that's the same open question as above. But we do know that *if* function-preserving substitution is possible, the resulting system will make just the same claims, the same behavior, and so on, as the original system. In fact we can know, in advance, precisely how the system will look from the third-person point of view. And even from the first-person point of view, I know that *if* some of my neurons are switched with identically-functioning silicon chips, I will come out swearing up and down that my qualia never changed. So in the relevant sense, I think we already know as much as we will ever know about how such a system will be.

Indeed, I think that if such substitution is ever possible, nobody will doubt the invariance principle for long. All it will take is a couple of substitutions, with subjects asserting that nothing has changed, and we will hear that there is "empirical evidence" that function-preserving substitution preserves conscious experience. The conclusion may be disputed by a handful of skeptical philosophers, but the subject's own word will be hard to resist. So I think that even now, the conditional assertion - *if* a functional isomorph is possible, *then* it will have the same sort of conscious experience - is at least as safe a bet.

Lowe thinks that the invariance principle "sells out completely" to functionalism, but this is a misunderstanding. Even many dualists hold that two subjects with the same brain state will have the same conscious state; presumably they are not thereby "selling out to physicalism", except in a highly attenuated sense of the latter. Consciousness is not reduced to a physical state; it is merely associated with one. By the same measure, to hold that two subjects in the same functional state have the same conscious state is not to sell out to functionalism, except in an attenuated sense. Consciousness is not reduced to a functional state; it is merely associated with one. Functional states, like physical states, turn out to determine conscious states with natural but not logical necessity. The resulting position, nonreductive functionalism, is compatible with the rich construal of mentality that reductive functionalism tacitly denies, precisely because a *logical* connection between function and experience is avoided.

Lowe may think that even a nonreductive functionalism is a bad thing, but to make that case, further reasons are required. For my part, I think that nonreductive functionalism stands a chance of capturing the most plausible and attractive elements of functionalist doctrines, while ignoring their reductive excesses.

Seager finds it odd that there should be laws connecting complex functional organizations to experience. I think that he is right and wrong to be worried about this. It would indeed be very odd if there were *fundamental* laws connecting complex organizations to experience (just as it would be odd if there were fundamental laws about telephones), but I do not claim that such laws exist. The invariance principle is intended as a non-fundamental law: eventually it should be the consequence of more fundamental laws that underlie it. Such laws need not invoke complex functional organization directly; they might instead invoke some simple underlying feature, such as information. As long as this feature is itself an organizational invariant (as information plausibly is), the invariance principle may be a consequence.

Seager also worries about the fineness of organizational grain required to duplicate experience. I discuss this in my book: the grain needed for the fading and dancing qualia argument to go through is one that is sufficiently fine to capture the mechanisms that support our behavioral dispositions, such as our dispositions to make certain claims, and also sufficiently fine to allow either (a) that any two realizations be connected by a near-continuous spectrum of realizations (for the fading qualia argument), or (b) that any two realizations be connected by a chain of realizations such that neighboring links in the chain differ only over a small region (for the dancing qualia argument). It is not impossible that a less fine grain will also suffice to duplicate experience, but the arguments will give no purchase on these cases. Seager worries that nature does not know about levels of organization, but again this would be a worry only if the invariance principle were held to be a fundamental law.

Finally, Seager thinks that the association of experience with functional organization leads to a particularly worrisome form of epiphenomenalism. I think it is clear, however, that the arguments he invokes apply to *any* association of experience with physical properties. There is indeed an interesting problem of "explanatory exclusion" to worry about, as I discussed above, but nothing about this problem is specific to the invariance principle or to any of the psychophysical laws I propose.

(3) The double-aspect analysis of information is by far the most speculative and tentative part of my article, and it is surely the most likely to be wrong. Indeed, as I say in my book, I think it is more likely than not to be wrong, but I put it forward in the hope that it might help progress toward a more satisfactory theory. So I am far from sure that I can defend it against every possible criticism. That being said, I think that a couple of the criticisms of the information-based approach may rest on misinterpretations.

Lowe resists my invocation of Shannonian information as "inappropriate for characterizing the cognitive states of human beings." But as before, I am not trying to *reduce* mental states to information processing. Such processing is instead invoked as a potential key to the physical *basis* of consciousness. True, the double-aspect view implies that consciousness has formal properties that mirror the formal properties of

the underlying information; I think this claim is clearly plausible from phenomenological investigation, but it is nowhere claimed that these formal properties exhaust the properties of consciousness. Just because the skeletal framework is syntactic, for example, there is nothing to prevent irreducible non-syntactic properties from being present as well. In fact, it is obvious that there are phenomenal properties over and above these formal properties: such properties are precisely what make the phenomenal realization of the information so different from the physical realization. Shannonian information at best provides a framework around which a theory of these intrinsic properties can be hung.

Varela is similarly "dumbfounded" by my appeal to this sort of information, because of the "outmoded cybernetic tradition" it invokes. I am not nearly as certain as Varela that Shannonian information (as opposed to the cyberneticist use of it) is outmoded; indeed, I think one can argue that information states of the kind I describe in my book play a central role even in the computationalist, connectionist, and "embodied" frameworks that Varela endorses. These frameworks may add something to information states - such as a semantic content, or a context within the world - but all these frameworks invoke certain "difference structures" and their causal roles in a cognitive system. And precisely because this difference structure captures an important formal isomorphism between aspects of conscious states and the underlying physical states, the concept of information may provide a framework within which we can make progress. Once it is clear that experience is not being *reduced* to information, I think the way is cleared for information to play a useful formal role, and perhaps even to play a role in the underlying metaphysics.

Libet, **Hardcastle**, and **Velmans** note that some information is nonconscious. As I discuss in my book, there are two ways to deal with this. The first is to find further constraints on the sort of information that is associated with experience; it is entirely possible that some such constraint may play a role in the psychophysical laws. (Velmans offers some interesting suggestions about such constraints, although none of them seem likely candidates to be truly fundamental.) The other possibility is to accept that all information has an experiential aspect: while not all information is realized in *my* consciousness, all information is realized in *some* consciousness. This is counterintuitive to many, but I do not think the possibility can be immediately dismissed. I will discuss it when I discuss panpsychism below.

The ontology underlying the informational picture (which **Velmans** worries about) remains open. I discuss a number of possible interpretations of it in my book. I am most attracted to a Russellian interpretation on which experience forms the "intrinsic" (or realizing) aspect of informational states which are fundamental to physics but characterized by physics only extrinsically. There is at least a kinship between the informational model and the Russellian metaphysics here, and exploiting it would lead to definite double-aspect ontology. ("Physics is information from the outside; experience is information from the inside.") But I am not certain that this can be made to work, and more straightforwardly dualistic interpretations are also available.

I favor the informational view largely because when I look for regularities between experience and the physical processes that underlie it, the most striking correspondences all lie at the level of information structures. We have to find *something* in underlying physical processes to link experience to, and information seems a plausible and universal candidate. Perhaps the biggest concern about this view is that

these informational structures do not lie at a *fundamental* level in physical processes; as **Bilodeau** notes, they are curiously abstract to play a role in a fundamental theory. On the other hand, there are ways of seeing information as fundamental to physics itself, so there may be ways in which a connection at a fundamental level can be leveraged to support this striking connection at the macroscopic level. But all that is very much in the realm of open questions.

4 POSITIVE PROPOSALS

A number of contributors made positive proposals about how the hard problem might be approached. These divide into (1) neuroscientific and cognitive approaches; (2) phenomenological approaches; (3) physics-based approaches, and (4) fundamental psychophysical theories. I will not try to assess each proposal at great length, but I will say a few words about the approaches and their relationship to my framework.

4.1 Neuroscientific and cognitive approaches

Proposals with a neurobiological and cognitive flavor were made by **Crick and Koch**, **Baars**, and **MacLennan**. The philosophical orientations of these range from reductionism to property dualism; this alone illustrates that a neurobiological approach to consciousness is compatible with many different philosophical views. Even if neurobiology and cognitive science *alone* cannot solve the hard problem, they may still play a central role in developing a theory.

Crick and Koch come closest to a reductionist view, although they are appropriately tentative about it. They first divide up the hard problem into three parts and offer an interesting solution to the third, concerning the incommunicability of experience. I think their idea here - that only relations are communicable because only relations are preserved throughout processing - is largely correct. That is, all that is communicable are differences that make a difference, or information states. Of course this is strictly speaking one of the "easy" problems, but it clearly has a close connection to the hard problem; I expect that a good cognitive account of what we can and cannot communicate about consciousness will lead to some very useful insights about the hard problem itself. I develop this point and tie it to an informational view of consciousness in Chapter 8 of my book.

On the hard problem, Crick and Koch suggest that it may be promising to focus first on "meaning". I am less sure about this: meaning seems to be almost as difficult a concept as consciousness, and perhaps even more ambiguous. If one invokes a purely functional construal of meaning - so that meaning comes down to certain correlations with the environment and certain effects on later processing - then a neurobiological account of meaning may be forthcoming, but such a functional account will not tell us why the meaning should be consciously experienced. And if one invokes a richer construal of meaning - one on which meaning is more closely tied to consciousness, for example - then there is more chance that an account of meaning may yield an account of consciousness, but a functional explanation of meaning becomes much less likely. Nevertheless I imagine there are useful insights to be had by treading this path, whether or not it leads to a solution to the hard problem.

An intermediate line is taken by **Baars**, who argues that a functional theory can at least shed considerable light on subjective experience, but who does not claim that it solves the hard problem. Indeed, he thinks the hard problem is too hard to be solved for now, because it involves an implausible criterion. I think that Baars misinterprets the hard problem slightly, however. To solve the hard problem we need not actually evoke all relevant experiences in ourselves (his "empathy criterion"). The point is not to *experience* what it is like to be a bat (although that would be nice!), but rather to explain why there is anything it is like to be a bat or a human at all. And this seems like a perfectly reasonable scientific question.

Baars also notes that there are deep causal connections between "easy" and "hard" aspects of our mental lives. This is certainly correct; indeed, I pointed out some such connections in my article. There seems to be a tight connection between global availability and consciousness, for example, as Baars suggests. So this sort of connection is quite compatible with my framework: the distinction between the easy and hard problems is a *conceptual* distinction, not a claim that the two have nothing to do with each other.

In particular, even once these causal interconnections are granted, one can still ask how and why the "easy" aspects are tied to the "hard". In conversation, Baars has suggested that one should just regard this as a brute fact, noting that psychologists are used to dealing in brute facts! So one might just take it as a brute fact that the contents of a global workspace are consciously experienced, for example. I think there is something to this, but one has to note that this brute fact has some strong consequences. For a start, it implies that a theory of consciousness requires explanatorily primitive principles over and above the facts about processing. Even if "easy" and "hard" phenomena are two different aspects of the same thing, as Baars suggests, this still requires some further principle to tie the two aspects together, and indeed to explain why there are two aspects in the first place.

Of course it is most unlikely that the whole problem will be solved in one bite, so it is entirely reasonable for Baars to leave things at the level of a connection between the global workspace and consciousness. This reflects a common strategy for dealing with consciousness in those areas of psychology that take it seriously: take the existence of consciousness for granted, and investigate just how and where it maps onto cognitive processing. (The literature on the properties of conscious vs. unconscious processes can be read this way, for example.) This way the roots of consciousness may be located, and the path may be cleared for a theory of the underlying connection.

MacLennan aims to take the next step, searching for a simple theory that explains the connection. He accepts that there is an irreducible phenomenal aspect that is systematically associated with neural processes, yielding a property dualism similar to mine but with a neurodynamical flavor; and he develops some ideas about the "deep structure" of the link between neural processes and experience.

I think MacLennan's idea of "protophenomena" (or "phenomenisca") as basic elements of consciousness is particularly interesting, and promises considerable rewards if it can be further developed. For a precise theory, I think we will need an account of (a) precisely when a protophenomenon is associated with a physical process, (b) what sort of protophenomena will be associated, depending on the characteristics of the physical process, and (c) the principles by which protophenomena combine into a unified conscious

experience.

None of these questions are trivial, although MacLennan makes a start on all of them. His answer to (a) relies on a one-activity-site-one-protophenomenon principle; for my part I would be surprised if things were so straightforward. It might be that protophenomena are determined by informational states of the system that are not straightforwardly localized, for example. He does not have too much to say about (b) - precisely what makes for the difference between visual and auditory protophenomena, for example? - but he has a preliminary analysis of (c). I suspect that (c) (an analog of the problems faced by the Russellian metaphysics described earlier) may turn out to be the hardest question of all.

In any case, I see the central parts of the projects of Crick and Koch, Baars, and MacLennan as all being compatible with the research program I envisage on the hard problem. At the nuts-and-bolts level, we must try to isolate the neural processes associated with consciousness, and to find detailed and systematic associations between these processes and characteristics of conscious experience. We should do the same at a cognitive level, where it may be that we will find "cleaner" associations if less detail, along with a way of integrating key elements of the neural story into a big picture. A clean association between global availability and consciousness, for example, promises to help make sense of messier associations involving various specific neural processes. Finally, we should search for the fundamental principles that underlie and explain these associations, boiling things down to as simple a system as possible.

All this is compatible both with the scientific worldview and with the irreducibility of consciousness. Once released from the insistent tug of the reductive dream, we are free to engage in the project of relating consciousness to physical processes on its own terms. The resulting science may be all the richer for it.

4.2 Phenomenological approaches

Shear and **Varela** concentrate on phenomenological approaches to the hard problem. I think that such an approach must be absolutely central to an adequate science of consciousness: after all, it is our own phenomenology that provides the data that need to be explained! If we are to have a detailed psychophysical theory, as opposed to a mere ontology, then we will have to catalog and systematize these data much as happens elsewhere in science; and to do this, patient attention to one's own experience is required.

Of course there are deep methodological problems here. The first is the old problem that the mere act of attention to one's experience transforms that experience. As we become more patient and careful, we may find that we are studying data that are transformed in subtle ways. This is not too much of a problem at the start of investigation -- we have a long way to go until this degree of subtlety even comes into play - but it may eventually lead to deep paradoxes of observership. Phenomenologists from both East and West have proposed ways to deal with this problem, but I think it has a certain resilience. Even if there do turn out to be limits on the fineness of this method's grain, however, I have no doubt that coarse-grained methods can take us a long way.

The second problem is that of developing a language - or better, a formalism - in which phenomenological data can be expressed. In other areas, the advent of such formalisms has led to rapid progress. We still seem to be far from such a formalism here, however. The notorious "ineffability" of conscious experience plays a role here: the language we have for describing experiences is largely derivative on the language we have for describing the external world. Perhaps, as Thomas Nagel has suggested, the structural properties of experience (e.g., the geometric structure of a visual field) will be most amenable to the possibility of formal expression, whether in informational, geometric, or topological terms, or in other terms entirely. I suspect that the residual non-structural properties will pose special problems.

The third difficulty lies in the failure, or at least the limitations, of incorrigibility: our judgments about experience can be wrong. I don't think this difficulty is as damning for phenomenology as it is sometimes made out to be; after all, our judgments about external data can be wrong, too, but science manages just fine. What is important is that our judgments about experience are accurate by and large, particularly when we are paying careful and patient attention. Our introspection must also be *critical*: we must take care to consider any ways in which it might be going wrong. But if our phenomenological judgments pass these tests, I think one is justified in taking them to be reliable.

Shear's and Varela's papers together make a strong case that a sophisticated phenomenological study is possible. In Shear's wide-ranging paper, the remarks about "pure consciousness" are particularly intriguing. I confess that I find myself among the skeptics where this notion is concerned. I am not sure that I can imagine a consciousness without quality: would not even a "void" experience have a certain voidish quality? (Shear's own position is appropriately cautious here.) But perhaps this is only because I have never experienced such a thing myself. The idea is appealing, at any rate, in the same sort of way that the Russellian idea of a physical world without intrinsic qualities is appealing: the appeal manifests itself both in spite of and because of its flirtation with incoherence. And the potential link that Shear suggests between this idea and a fundamental theory is certainly suggestive.

I am also sympathetic with much of Varela's discussion, in its shape if not in every detail. Varela takes himself to differ with me on some central points, but I am not sure why. The main difference between us seems to be one of emphasis: he emphasizes the phenomenological data, whereas I emphasize the systematicity in the relationship between these data and underlying processes. Perhaps he takes my "extra ingredient" or "theoretical fix" to be something more reductive than I intended. Varela himself seems to endorse the need for an extra ingredient in our theories - namely experience itself - which fits my program well. He may differ by doubting the likelihood of simple underlying laws connecting the physical and phenomenal domains; but if so, he does not give his reasons in this article. In any case, the idea of "neurophenomenology" sounds eminently sensible to me. The test will be whether it can be cashed out in the form of detailed results.

It would be overambitious to suppose that phenomenology by itself offers a solution to the hard problem. The ontological debates are as hard as ever, and phenomenology is largely neutral on them (except, perhaps, in rejecting type-A materialism). But it is absolutely central to the *epistemology* of the hard

problem: without it, we would not even know what needs explaining. In most areas of science, we need an adequate epistemology to get a detailed theory off the ground, and there is no reason to suspect that the case of consciousness will be any different. If so, the sort of careful study advocated by Shear and Varela will be a central component in the path to a solution.

4.3 Physics-based approaches

In getting an empirical theory of consciousness off the ground, the two areas just discussed will play the central roles. Neuro/cognitive science will provide the third-person data and phenomenology will provide the first-person data. As all this goes on, theorists of all stripes will seek to systematize the connection between the two. In the early stages, this connection will be strongest at the "surface" level: researchers will isolate correlations between fairly complex neuro/cognitive processes and relatively familiar characteristics of conscious experience. This high-level project may well be the solid core of consciousness research for many years to come. As the project develops, though, there will be an increasing drive to find the deep structure that underlies and explains these high-level connections, with the ultimate goal being a fundamental psychophysical theory.

We are not close to having such a fundamental theory yet, but this need not stop us from speculating about its form. Many contributors to this symposium do just that, offering proposals about links between consciousness and physical processes at the most fundamental level. In this section and the next, I will discuss these proposals. Those with conservative tastes might stop here: what follows is largely untrammelled speculation in physics and metaphysics about what may be required to bring consciousness within the natural order. I do not know whether any of this is on the right track, but there are plenty of interesting ideas with which I am more than happy to play along.

A number of contributors suggest approaches in which physics plays a central role. I expressed some criticism of physics-based proposals in the keynote paper, but mostly insofar as these were offered as *reductive* explanations of consciousness. ("Neurons can't do the job, but quantum mechanics can."). None of the current contributors offer that sort of account. Most of them instead offer proposals on which consciousness is taken as fundamental, and is related nonreductively to the entities in physical theories, perhaps in the hopes of finding a natural place for consciousness in the natural order. Such suggestions are not subject to the same sort of criticism, and they certainly cannot be ruled out *a priori*.

The difference between the two sorts of physics-based proposals is most apparent in the article by **Hameroff and Penrose**. Previous work had given me the impression that their aim was to explain consciousness wholly in terms of quantum action in microtubules; but this paper makes it explicit that consciousness is instead to be taken as fundamental. In essence, Hameroff and Penrose offer a psychophysical theory, postulating that certain quantum-mechanical reductions of the wave function, brought on when a certain gravitational threshold is attained, are each associated with a simple event of experience. They suggest a kinship with Whitehead's metaphysics; the view might also fit comfortably into the Russellian framework outlined earlier.

This is an intriguing and ambitious suggestion. Of course the details are a little sketchy: after their initial postulate, Hameroff and Penrose concentrate mostly on the physics of reduction and its functioning in microtubules, and leave questions about the explanation of experience to one side. Eventually it would be nice to see a proposal about the precise form of the psychophysical laws in this framework, and also to see how these billions of microscopic events of experience might somehow yield the remarkable structural properties of the single complex consciousness that we all possess. I am cautious about this sort of quantum-mechanical account myself, partly because it is not yet clear to me that quantum mechanics is essential to neural information-processing, and partly because it is not easy to see how quantum-level structure corresponds to the structure one finds in consciousness. But it is not impossible that a theory might address these problems. To know for sure, we will need a detailed explanatory bridge.

Stapp offers a very different sort of quantum-mechanical proposal. Instead of trying to constitute experience out of many low-level quantum-mechanical events, he takes consciousness as a given, and offers a theory of the role it plays in collapsing physical wave functions, thus showing how it might have an impact on the physical world. As I said earlier, this sort of "collapse" interpretation of quantum mechanics needs to be taken very seriously - in the interests both of giving a good account of quantum mechanics and of giving a good account of consciousness - and Stapp's, as developed in a number of papers, is perhaps the most sophisticated version of such an interpretation to date. It certainly offers the most natural picture in which consciousness plays a role in influencing a non-causally-closed physical world.

Stapp's paper is neutral on some central questions that a theory of consciousness needs to answer. He says quite a lot about his mental-to-physical laws, characterizing the role of consciousness in wave-function collapse, but he does not say much about the physical-to-mental laws which will presumably be at the heart of a theory. Such laws will tell us just *which* physical processes are associated with consciousness, and what *sort* of conscious experience will be associated with a given physical process. (Of course, we know that experiences have "actualizations" as a physical correlate; but given that Stapp wants pre-existing experiences to *cause* the actualizations, we need some independent physical criterion for experience. This would then yield a physical criterion for actualization in turn.) As it stands, Stapp's picture seems compatible with almost any physical-to-mental laws. Stapp offers some suggestions about such laws in his book (Stapp 1993), where he proposes that experience goes along with "top-level processes" in the brain; but perhaps it is a virtue of Stapp's broader proposal about the causal role of consciousness that many different psychophysical theories can benefit by invoking it.[*]

*[[[One intriguing if far-out possibility: if Stapp's proposal were granted, it might even be that experimental physics could help determine the psychophysical laws, and determine which systems are conscious, at least in principle. It turns out that different proposals about the physical criteria for collapse have subtly different empirical consequences, although they are consequences that are practically impossible to test in general (see Albert 1992 for discussion). So at least in principle, if not in practice, one could test for the presence or absence of collapse in a given system, and thus for the presence or absence of experience!]]]

Clarke suggests a different connection between physics and consciousness, rooted in the nonlocality of both. The nonlocality of the former is less controversial, in a way: nonlocal causal influences are present in most interpretations of quantum mechanics, with the exception of those by Everett (1973) and Cramer

(1986), and nonlocal constitution of physical states is present on most of these in turn. The sense in which mind is nonlocal is less clear to me. I am sympathetic with Clarke's point that mind is not located in physical space, but I am not sure of the link between these two sorts of nonlocality. Clarke argues that the physical structure that supports mind has to be nonlocal; but all that is clear to me is that it has to be *nonlocalized*, or distributed across space, which is equally possible on a classical theory. But perhaps nonlocal constitution of a physical state could be linked to the unity of consciousness, especially on a view which identifies consciousness with a physical state in such a way that unified consciousness requires a unified substrate: nonlocal physical constitution might unify the basis? The idea might also help in a Russellian metaphysics, though I am not sure that it is required.

Another appeal to physics is made by **McGinn**, who suggests that accommodating consciousness within the natural order will require a radically revised theory of space. A question immediately suggests itself: will this theory be forced on us to explain (third-person) empirical evidence, or just to accommodate consciousness? I suspect that it must be the latter. All sorts of revisions in our physical theories are made to explain the external world, but they always leave theories cast in terms of some basic mathematical structures and dynamics (whether Euclidean space, four-dimensional space-time, or infinite-dimensional Hilbert space). There are principled reasons why structure and dynamics is all we could possibly need to explain external evidence; and given *any* theory cast solely in terms of structure and dynamics, the further question of consciousness will arise.

So it seems to me that McGinn needs an empirically adequate theory of space to be revised or supplemented in some fundamental way to accommodate consciousness, while leaving its external predictions intact. But McGinn also strongly wants to avoid epiphenomenalism (see McGinn 1996). I think that the natural way (perhaps the only way) to satisfy these requirements is along the Russellian lines suggested above: there is a pervasive intrinsic property of physical reality, a property which *carries* the structure and dynamics specified in physical theory but is nevertheless not revealed directly by empirical investigation, and which enables the existence of consciousness. This picture seems to square well with McGinn's remarks about a "hidden dimension" of physical reality. Concentrating on space in particular, we might perhaps think of this property as the "medium" in which the mathematical structures of space are embedded.

It seems clear, at any rate, that McGinn's "hidden dimension" requires us to postulate *something* new and fundamental over and above what is empirically adequate. As such it seems that he is embracing option (2) of the dilemma I posed him earlier in this paper. And this new fundamental property is a sort of "proto-experience", at least in the sense that it enables the existence of experience. If so then McGinn's view, when unpacked, is in the same sort of ballpark as the views I am advocating. Of course McGinn *could* be right that we will never be able to form such a theory, for example because of our inability to grasp the relevant proto-experiential concept. On the other hand, he could be wrong; so I for one will keep trying.

Bilodeau takes the most radical physics-based approach, holding (I think) that we have to abandon the idea that there are objectively existing states in fundamental physics. Instead, physical reality crystallizes in some way as a product of experience and the process of inquiry. Once we see that experience is fundamental to the very nature of physical reality in this way, the hard problem may go away.

Bilodeau suggests that this picture is the most natural upshot of quantum mechanics, appealing especially to the writings of Bohr. Now I think this picture is certainly not *forced* on us by quantum mechanics - there are plenty of ways of making sense of quantum mechanics while maintaining the idea that fundamental physical reality has an objective existence, if only in the form of a superposed wave function. Bilodeau clearly finds these interpretations unappealing, but I (like many others) find them much more comprehensible. Given that macroscopic physical reality has an objective existence, it seems that its causal antecedents must have objective existence (otherwise why would it come into existence?), and in the process of explanation we are relentlessly driven to causal antecedents at more and more fundamental levels. So the only way I can make sense of the idea that fundamental physical reality does not have objective existence is as a form of idealism, on which *all* physical reality is present only within experience. Bilodeau disclaims this interpretation, however, so this may be a cognitive limitation on my part.

In any case it seems that even under Bilodeau's reasoning, there still needs to be an explanatory theory connecting experiences and brain processes. I am not quite sure what the shape of such a theory will be, but perhaps his version of the metaphysics will be able to give a natural version of it. It would be very interesting to see some of the details.

4.4 Fundamental psychophysical theories

Some of the most intriguing pieces, to me, are those that speculate about the shape of a fundamental theory of consciousness. Many of these proposals invoke some form of panpsychism. Panpsychism is not *required* for a fundamental theory; it is not written in stone that fundamental properties have to be ubiquitous. **Libet** and **Stapp**, for example, both invoke fundamental theories without invoking panpsychism. But the idea of a fundamental theory certainly fits well with panpsychism, and the proposals by **Hut and Shepard**, **Rosenberg**, and **Seager** are all explicitly panpsychist.

Some contributors (e.g. **Mills** and **Hardcastle**) roll their eyes at the idea of panpsychism, but explicit arguments against it are surprisingly hard to find. Rosenberg and Seager give nice defenses of panpsychism against various objections. Indeed, both upbraid me for not being panpsychist *enough*. I do not know whether panpsychism is true, but I find it an intriguing view, and in my book I argue that it deserves attention. If a simple and powerful predictive theory of consciousness ends up endorsing panpsychism, then I do not see why we should not accept it.

Panpsychist views need not ascribe *much* of a mind to simple entities. Sometimes the term "panexperientialism" is used instead, to suggest that all that is being ascribed is some sort of experience (not thought, not intelligence, not self-awareness), and a particularly simple form of experience at that. And some versions do not even go this far. Instead of suggesting that experience is ubiquitous, such views suggest that some *other* property is ubiquitous, where instantiations of this property somehow jointly constitute experience in more complex systems. Such a property might be thought of as a *proto-experiential* property, and the associated view might more accurately be thought of as *panprotopsychism*.

Of course it is very hard to form a conception of protoexperiential properties. We know no set of *physical* properties can constitute experience, for familiar reasons. But perhaps some quite alien property might do the job. I was particularly intrigued by **Hut and Shepard's** postulation of a property 'X', where X stands to consciousness as time stands to motion. That is, just as time enables the existence of motion, in combination with space, X enables the existence of consciousness, in combination with the basic dimensions of space-time. This offers an elegant picture of proto-experience quite different from the tempting picture on which proto-experience is "just like experience but less so".

In a way, Hut and Shepard's proposal has a lot in common with **McGinn's** suggestion of a "hidden dimension" of space which enables the existence of consciousness. As with McGinn, once can ask whether the dimension is truly "hidden", or whether it will manifest itself in our external observations (the physics we have now does a pretty good job, after all). As before, I suspect that such a property *has* to be hidden, as an empirically adequate theory can always be cast in terms of structure and dynamics that are compatible with the absence of experience. Thus, as before it seems that the new dimension will either (a) be epiphenomenal to the other dimensions (or at least to the projections of those dimensions that we have access to), or (b) related to them as a sort of Russellian "realizing" property, carrying the structure in one of these dimensions and making it real. The latter would be particularly compatible with the idea of turning the hard problem "upside down", on which physical reality is itself somehow derivative on underlying (proto)experiences.

Rosenberg offers a detailed defense of panpsychism, and makes a number of points with which I am particularly sympathetic. He makes a strong case against the existence of fundamental laws that connect consciousness to mere complexity, to aspects of functioning, or to biological properties. While I think there is nothing *wrong* with the idea of a nonpanpsychist fundamental theory, Rosenberg's discussion eliminates some of the most obvious candidates. (Another possibility worth considering, though: several simple laws might *combine* to imply that experience only comes into existence in certain complex cases.) And he begins to unpack what panpsychism might involve in a way that makes it clear that the idea is at least coherent.

Rosenberg also makes a strong case for an *integrated* view of nature, on which consciousness is not a mere tacked-on extra. My keynote paper may carry a flavor of the latter (except for the final paragraph of section VII), but I think the former is the ultimate goal. Perhaps the best path to such an integrated view is offered by the Russellian picture on which (proto)experiential properties constitute the intrinsic nature of physical reality. Such a picture is most naturally associated with some form of panpsychism. The resulting integration may be panpsychism's greatest theoretical benefit.

Seager also provides some motivation for panpsychism, and gives a particularly interesting accounting of its problems. I think his "completeness problem" (a version of the epiphenomenalism problem) is mitigated by embracing the Russellian interpretation, on which the fundamental (proto)experiences are part of the causal order, although there will always be residual worries about *explanatory* superfluity. (Giving experiences certain anomalous effects doesn't help here; experience-free structural explanations are just as possible either way.) This view would also solve his "no-sign" problem: we cannot expect to have external access to the intrinsic properties that underlie physical dispositions. A solution to the "not-

mental" problem must likely wait until we *have* a theory; presumably we will then be justified in attributing (proto-) mentality in certain cases precisely because of the theory's indirect explanatory benefits in explaining our *own* experiences. A version of the "unconscious mentality" problem will apply to any view that postulates proto-experiential rather than experiential properties at the fundamental level (how does experience emerge from non-experience?), but this need not be quite as hard as the original hard problem. We know that physical properties cannot imply experience, because of the character of physics, but novel intrinsic proto-experiential properties cannot be ruled out in the same way.

This leaves the "combination problem", which is surely the hardest. This is the problem of how low-level proto-experiential and other properties somehow together *constitute* our complex unified conscious experiences. (One might also think of it as the "constitution problem", to avoid the implication that constitution must work by simple combination; consider Hut and Shepard's non-combinatorial proposal, for example.) The problem could be bypassed altogether by suggesting that complex experiences are *not* constituted by the micro-experiences, but rather arise autonomously. This would hold true under many psychophysical theories, including some versions of an informational theory; its main disadvantage is that it once again threatens epiphenomenalism. To make experience causally relevant in the Russellian way, it seems that it *has* to be constituted out of the intrinsic natures of the fundamental causally relevant entities in physical theory. Unless we embrace an interactionist picture like **Stapp's** where there is fundamental causation at a high level, it seems that integrating experience into the causal order leads inevitably to the combination problem.

To solve the problem, we have to investigate the principles of composition to which experience is subject. The "problem" may well arise from thinking of experiential composition along the lines of physical composition, when it might well work quite differently. I suggest in my book, for example, that something more like *informational* composition might be more appropriate. Alternatively, we may try to keep a closer isomorphism between experiential composition and physical composition, but investigate nonstandard manners of physical composition. Seager's invocation of quantum coherence is an intriguing example of such a strategy: in this case physical composition yields a unity that might mirror the unity of experience. To the best of my knowledge, the evidence for widespread stable quantum coherent states at a macroscopic level in the brain is not strong, but this is nevertheless a strategy to keep in mind. A related quantum-mechanical strategy is discussed by Lockwood (1992), who also provides an illuminating discussion of the problem in general. There may well be other interesting ideas waiting to be explored in addressing the problem; it is likely to be a fruitful area for further inquiry.

Of course everything in these last two sections has the air of something put together in the metaphysical laboratory, to use Seager's phrase. It is all extraordinarily speculative, and has to be taken with a very large grain of salt. Like my own speculations about information, these suggestions have not yet been remotely developed to the point where they can be given a proper assessment - indeed, their largely undefined nature may be the reason that I am able to speak reasonably warmly of all of them! And most of them have not yet begun to provide a detailed explanatory bridge from the fundamental level to the complex experiences we know and love. I favor the informational view partly because it seems closer to providing such a bridge than proposals based directly in physics or elsewhere, but even this view is very sketchy in crucial places.

To have a fundamental theory that we can truly assess, we will need a fundamental theory with *details*. That is, we will need specific proposals about psychophysical laws, and specific proposals about how these laws combine, if necessary, so that ultimately we will be able to (1) take the physical facts about a given system, (2) apply the psychophysical theory to these facts, and thus (3) *derive* a precise characterization of the associated experiences that the theory predicts. As yet, we do not have a single theory that allows this sort of derivation. Indeed, as I noted above, we may first need to develop a proper formalism (informational, geometrical, topological?) for characterizing experiences before this project can get off the ground. And once we have such a formalism, it may well be extremely hard to devise a theory that even gives the right results in the simplest familiar cases. Once we *do* have a detailed theory that gives approximately correct results in familiar cases, however, we will know we are on the right track. The ultimate goal is a simple theory that gets things exactly right.

I imagine that it may well be many years until we have a good detailed theory. We will probably first have to concentrate on understanding the "macroscopic" regularities between processing and experience, and gradually work our way down to the fundamental principles that underlie and explain these regularities. Most researchers are now working at the macroscopic level, insofar as they are working on experience at all, and this is as it should be. But we can at least speculate about the form of a fundamental theory, in our more philosophical moments, and there is no reason why we should not try to come up with some details. Perhaps we will prove to have been terribly premature, but we will not know until we try. And in the meantime, I am sure that the attempt will be enlightening.

5 CONCLUSION

Taking a broad view of the metaphysics of the hard problem, here is the lay of the land as I see it.

(1) The first "choice point" is the question of whether there is a problem of consciousness at all, distinct from the problem of explaining functions. Some, the type-A materialists, deny this, though we have seen that there seem to be few good arguments for such a counterintuitive conclusion. Given that there *is* a further phenomenon that needs explaining, we have seen that one is forced to the conclusion that no reductive explanation of consciousness can be given, and that explanatorily primitive bridging principles are required.

(2) In a second choice point, some (the type-B materialists) try to preserve materialism by arguing that these principles are "identities". But we have seen that these explanatorily primitive identities are unparalleled elsewhere in science, are philosophically problematic, and require the invocation of a new and ungrounded form of necessity. In any case, the *form* of a theory of this sort will be just like the form of a theory that takes consciousness as fundamental, and these "identities" will function in our explanations just like fundamental laws.

(3) All other theories take experience (or proto-experience) as irreducible, along with irreducible principles relating it to the physical domain. The next choice point is whether to hold onto the causal closure of the physical. Denying this, perhaps through an invocation of wavefunction collapse in quantum

mechanics, leads to an interactionist dualism. But the advantages of this denial can be questioned.

(4) Given that the physical domain is a closed causal network, the next choice is that between views which put experience *outside* this network, with psychophysical laws that make experience epiphenomenal, or which put experience *inside* this network, by virtue of a Russellian monism on which the intrinsic properties of matter are proto-experiential. The latter offers the most attractive and integrated view, if the "combination problem" can be solved.

(5) The final choice point turns on the *form* of the psychophysical laws in our theory. This is the meatiest question of all, and can be engaged by researchers in all fields: the earlier questions require some tolerance for metaphysics, but this question is more straightforwardly "scientific". Much work on this question will be independent of specific choices on questions (2)-(4), though some aspects of these choices may inform one's approach to this question at some point.

Progress on the hard problem will likely take place at two levels. On the philosophical level, there will be an ongoing clarification of the issues surrounding (1)-(4), and the arguments for and against the various options at the various choice points. For my part I think the case for introducing new irreducible properties is hard to resist, but the choice points at (3) and especially (4) are still open. On a more concrete level, there will be progress toward specific laws as in (5). A combination of experimental study, phenomenological investigation, and philosophical analysis will lead us to systematic principles bridging the domains, and eventually we hope to be led to the underlying fundamental laws. In this way we may eventually arrive at a truly satisfactory theory of conscious experience.

Bibliography

Albert, D. 1992. *Quantum Mechanics and Experience*. Cambridge, MA: Harvard University Press.

Baars, B.J. 1996. Understanding subjectivity: Global workspace theory and the resurrection of the observing self. *Journal of Consciousness Studies* 3:211-17.

Bilodeau, D.J. 1996. Physics, machines, and the hard problem. *Journal of Consciousness Studies* 3:386-401.

Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Churchland, P.M. 1996. The rediscovery of light. *Journal of Philosophy* 93:211-28.

Churchland, P.S. 1996. The hornswoggle problem. *Journal of Consciousness Studies* 3:402-8

Clark, T. 1995. Function and phenomenology: Closing the explanatory gap. *Journal of Consciousness Studies* 2:241-54.

- Clarke, C.J.S. 1995. The nonlocality of mind. *Journal of Consciousness Studies* 2:231-40.
- Crick, F. and Koch, C. 1995. Why neuroscience may be able to explain consciousness. *Scientific American* 273(6):84-85.
- Cramer, J.G. 1986. The transactional interpretation of quantum mechanics. *Review of Modern Physics* 58:647-87.
- Denett, D.C. 1996. Facing backwards on the problem of consciousness. *Journal of Consciousness Studies* 3:4-6.
- Everett, H. 1973. The theory of the universal wave function. In (B.S. de Witt & N. Graham, eds.) *The Many-Worlds Interpretation of Quantum Mechanics*. Princeton: Princeton University Press.
- Kohler, D. 1947. *Gestalt Psychology*. New York: Liveright Publishing Corporation.
- Hardcastle, V.G. 1996. The why of consciousness: A non-issue for materialists. *Journal of Consciousness Studies* 3:7-13.
- Hawking, S. 1988. *A Brief History of Time*. Bantam Books.
- Hodgson, D. 1996. The easy problems ain't so easy. *Journal of Consciousness Studies* 3:69-75.
- Hut, P. & Shepard, R. Turning the "hard problem" upside-down and sideways. *Journal of Consciousness Studies* 3:313-29.
- Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.
- Libet, B. 1996. Solutions to the hard problem of consciousness. *Journal of Consciousness Studies* 3:33-35.
- Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford: Blackwell.
- Lockwood, M. 1992. The grain problem. In (H. Robinson, ed.) *Objections to Physicalism*. Oxford: Oxford University Press.
- Lowe, E.J. 1995. There are no easy problems of consciousness. *Journal of Consciousness Studies* 2:266-71.
- MacLennan, B. 1996. The elements of consciousness and their neurodynamical correlates. *Journal of Consciousness Studies* 3:409-24.

- Maxwell, G. 1978. Rigid designators and mind-brain identity. In (C.W. Savage, ed.) *Perception and Cognition: Issues in the Foundations of Psychology* (Minnesota Studies in the Philosophy of Science, Vol. 9). Minneapolis: University of Minnesota Press.
- McGinn, C. 1989. Can we solve the mind-body problem? *Mind* 98:349-66. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).
- McGinn, C. 1995. Consciousness and space. *Journal of Consciousness Studies* 2:220-30.
- McGinn, C. 1996. Review of *The Conscious Mind*. Times Higher Educational Supplement, April 5 1996, pp. vii-ix.
- Mills, E. 1996. Interactionism and overdetermination. *American Philosophical Quarterly* 33:105-115.
- Mills, E. 1996. Giving up on the hard problem of consciousness. *Journal of Consciousness Studies* 3:26-32.
- Muller, G.E. 1896. Zur Psychophysik der Gesichtsempfindungen. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane* 10:1-82.
- O'Hara, K. & Scutt, T. 1996. There is no hard problem of consciousness. *Journal of Consciousness Studies* 3.
- Papineau, D. 1996. Review of *The Conscious Mind Times Literary Supplement* 4864 (June 21, 1996), pp. 3-4.
- Place, U.T. 1956. Is consciousness a brain process? *British Journal of Psychology* 47:44-50. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).
- Price, M.C. 1996. Should we expect to feel as if we understand consciousness? *Journal of Consciousness Studies* 3:303-12.
- Robinson, W.S. 1996. The hardness of the hard problem. *Journal of Consciousness Studies* 3:14-25.
- Rosenberg, G.H. 1996. Rethinking nature: A hard problem within the hard problem. *Journal of Consciousness Studies* 3:76-88.
- Russell, B. 1927. *The Analysis of Matter*. London: Kegan Paul.
- Seager, W. 1995. Consciousness, information, and panpsychism. *Journal of Consciousness Studies* 2:272-88.

- Shear, J. 1996. The hard problem: Closing the empirical gap. *Journal of Consciousness Studies* 3:54-68.
- Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:291-315. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).
- Shoemaker, S. 1990. First-person access. *Philosophical Perspectives* 4:187-214.
- Smart, J.J.C. 1959. Sensations and brain processes. *Philosophical Review* 68:141-56. Reprinted in (D. Rosenthal, ed) *The Nature of Mind* (Oxford University Press, 1990).
- Stapp, H. 1993. *Mind, Matter, and Quantum Mechanics*. Springer Verlag.
- Varela, F. 1995. Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies* 3.
- Velmans, M. 1995. The relation of consciousness to the material world. *Journal of Consciousness Studies* 2:255-65.
- Warner, R. 1996. Facing ourselves: Incorrigibility and the mind-body problem. *Journal of Consciousness Studies* 3:217-30.
- White, S. 1986. Curse of the qualia. *Synthese* 68:333-68.

Response to Papineau

This is a letter I sent to the *Times Literary Supplement* in response to David Papineau's review of my book *The Conscious Mind: In Search of a Fundamental Theory*. The letter was published in their issue of August 2, 1996, in a somewhat mangled form that made a few points hard to follow.

How do we explain the existence of consciousness in a physical world? In his review of my book *The Conscious Mind*, David Papineau defends an odd form of materialism, holding that brain states are conscious states, and that is that. To one who asks for an *explanation* of this fact - why does being in a brain state feel any way at all? - Papineau insists that we unask the question. The identity between brain states and conscious states is not something to be further explained, but rather is a sort of explanatory primitive.

This seems to make consciousness more mysterious than ever, except that Papineau puts a positive spin on his views by holding that identities are not the sort of thing one explains. But this is false, or at least misleading. The only explanatorily primitive principles elsewhere in science are the fundamental laws of physics: from these (plus boundary conditions), all else follows. Someone possessing the concepts of "Samuel Clemens" and "Mark Twain" (to use Papineau's example) but ignorant of their identity could *deduce* that identity from a specification of the microphysical facts, for example. But Papineau concedes that the brain-mind link cannot be deduced from microphysics; so this link is explanatorily primitive in a much stronger sense.

I think that such explanatory primitiveness is the mark of a fundamental law of nature, and so (in my book) go on to investigate a "fundamental theory" of the mind-brain link, analogous to fundamental theories in physics. Papineau chooses to call the primitive relation an "identity" rather than a law, thus saving materialism in letter if not in spirit. I suspect that many will find that the most important point has already been conceded; but in any case there are further technical reasons, discussed in the book, why this maneuver does not make sound philosophical sense.

In brief: Papineau's position commits him to a notion that I call "strong metaphysical necessity", in which there is little reason to believe (it is unlike any necessity found elsewhere in philosophy), and which suffers from all sorts of problems. He avoids this issue by misrepresentation and elision: he wrongly implies that my analysis assumes a "descriptive" view of reference, and he unaccountably ignores the treatment of strong metaphysical necessity, which is there to handle precisely his sort of "escape route". His further talk of "reference by simulation" and an "antipathetic fallacy" merely reiterates, rather than justifies, his commitment to explanatorily primitive identities and strong metaphysical necessities.

Technicalities aside, it is hard to see how Papineau's declaring the mind/brain relationship an "identity" gets him off the hook in providing a detailed theory of consciousness. We would still like a theory to predict just *how* various brain states will feel, whether it feels like anything to be a different system such

as a computer, and so on (unless Papineau implausibly denies that there is a fact of the matter here). And as elsewhere in science, one would like to systematize and simplify the relevant principles as far as possible. The result might well look like the fundamental theories I recommend, whether one talks of identities or laws.

Papineau's critique of my positive views consists in the observation that I am committed to epiphenomenalism - another misrepresentation. Although I don't share Papineau's phobia toward this view, I discuss at length an interpretation that avoids both epiphenomenalism and Cartesian dualism. This is an adaptation of the Russellian idea, recently advocated by Michael Lockwood, that the intrinsic qualities of experience can play the role of the intrinsic qualities of matter, on which physics is silent. This idea turns out to fit nicely with the information-based theories sketched in the constructive half of the book (which Papineau skips entirely). I think the choice between ontologies here is still open, but there are plenty of interesting options.

It would be terrific if the problem of consciousness could be removed by a magic bullet as simple as Papineau's, and life would be a lot easier if every opponent's argument could be shoehorned into a familiar stereotype. But deep problems and real arguments have a habit of popping straight back up again.

David Chalmers
Department of Philosophy
University of Arizona
Tucson, AZ 85721, USA.

Go to:

- [David Chalmers' home page](#)
- [The Conscious Mind web page](#)

Facing Up to the Problem of Consciousness

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

[This appeared in the *Journal of Consciousness Studies* in 1995. Also online is my response, "[Moving Forward on the Problem of Consciousness](#)", to 26 articles commenting on this paper. That paper elaborates and extends many of the ideas in this one.]

1 Introduction

Consciousness poses the most baffling problems in the science of the mind. There is nothing that we know more intimately than conscious experience, but there is nothing that is harder to explain. All sorts of mental phenomena have yielded to scientific investigation in recent years, but consciousness has stubbornly resisted. Many have tried to explain it, but the explanations always seem to fall short of the target. Some have been led to suppose that the problem is intractable, and that no good explanation can be given.

To make progress on the problem of consciousness, we have to confront it directly. In this paper, I first isolate the truly hard part of the problem, separating it from more tractable parts and giving an account of why it is so difficult to explain. I critique some recent work that uses reductive methods to address consciousness, and argue that such methods inevitably fail to come to grips with the hardest part of the problem. Once this failure is recognized, the door to further progress is opened. In the second half of the paper, I argue that if we move to a new kind of nonreductive explanation, a naturalistic account of consciousness can be given. I put forward my own candidate for such an account: a nonreductive theory based on principles of structural coherence and organizational invariance and a double-aspect view of information.

2 The easy problems and the hard problem

There is not just one problem of consciousness. "Consciousness" is an ambiguous term, referring to many different phenomena. Each of these phenomena needs to be explained, but some are easier to explain than others. At the start, it is useful to divide the associated problems of consciousness into "hard" and "easy" problems. The easy problems of consciousness are those that seem directly susceptible to the standard methods of cognitive science, whereby a phenomenon is explained in terms of computational or neural

mechanisms. The hard problems are those that seem to resist those methods.

The easy problems of consciousness include those of explaining the following phenomena:

- the ability to discriminate, categorize, and react to environmental stimuli;
- the integration of information by a cognitive system;
- the reportability of mental states;
- the ability of a system to access its own internal states;
- the focus of attention;
- the deliberate control of behavior;
- the difference between wakefulness and sleep.

All of these phenomena are associated with the notion of consciousness. For example, one sometimes says that a mental state is conscious when it is verbally reportable, or when it is internally accessible. Sometimes a system is said to be conscious of some information when it has the ability to react on the basis of that information, or, more strongly, when it attends to that information, or when it can integrate that information and exploit it in the sophisticated control of behavior. We sometimes say that an action is conscious precisely when it is deliberate. Often, we say that an organism is conscious as another way of saying that it is awake.

There is no real issue about whether *these* phenomena can be explained scientifically. All of them are straightforwardly vulnerable to explanation in terms of computational or neural mechanisms. To explain access and reportability, for example, we need only specify the mechanism by which information about internal states is retrieved and made available for verbal report. To explain the integration of information, we need only exhibit mechanisms by which information is brought together and exploited by later processes. For an account of sleep and wakefulness, an appropriate neurophysiological account of the processes responsible for organisms' contrasting behavior in those states will suffice. In each case, an appropriate cognitive or neurophysiological model can clearly do the explanatory work.

If these phenomena were all there was to consciousness, then consciousness would not be much of a problem. Although we do not yet have anything close to a complete explanation of these phenomena, we have a clear idea of how we might go about explaining them. This is why I call these problems the easy problems. Of course, "easy" is a relative term. Getting the details right will probably take a century or two of difficult empirical work. Still, there is every reason to believe that the methods of cognitive science and neuroscience will succeed.

The really hard problem of consciousness is the problem of *experience*. When we think and perceive, there is a whirl of information-processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is *something it is like* to be a conscious organism. This subjective aspect is experience. When we see, for example, we *experience* visual sensations: the felt quality of redness, the experience of dark and light, the quality of depth in a visual field. Other experiences go along with perception in different modalities: the sound of a clarinet, the smell of mothballs. Then there are bodily sensations, from pains to

orgasms; mental images that are conjured up internally; the felt quality of emotion, and the experience of a stream of conscious thought. What unites all of these states is that there is something it is like to be in them. All of them are states of experience.

It is undeniable that some organisms are subjects of experience. But the question of how it is that these systems are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does.

If any problem qualifies as *the* problem of consciousness, it is this one. In this central sense of "consciousness", an organism is conscious if there is something it is like to be that organism, and a mental state is conscious if there is something it is like to be in that state. Sometimes terms such as "phenomenal consciousness" and "qualia" are also used here, but I find it more natural to speak of "conscious experience" or simply "experience". Another useful way to avoid confusion (used by e.g. Newell 1990, Chalmers 1996) is to reserve the term "consciousness" for the phenomena of experience, using the less loaded term "awareness" for the more straightforward phenomena described earlier. If such a convention were widely adopted, communication would be much easier; as things stand, those who talk about "consciousness" are frequently talking past each other.

The ambiguity of the term "consciousness" is often exploited by both philosophers and scientists writing on the subject. It is common to see a paper on consciousness begin with an invocation of the mystery of consciousness, noting the strange intangibility and ineffability of subjectivity, and worrying that so far we have no theory of the phenomenon. Here, the topic is clearly the hard problem - the problem of experience. In the second half of the paper, the tone becomes more optimistic, and the author's own theory of consciousness is outlined. Upon examination, this theory turns out to be a theory of one of the more straightforward phenomena - of reportability, of introspective access, or whatever. At the close, the author declares that consciousness has turned out to be tractable after all, but the reader is left feeling like the victim of a bait-and-switch. The hard problem remains untouched.

3 Functional explanation

Why are the easy problems easy, and why is the hard problem hard? The easy problems are easy precisely because they concern the explanation of cognitive *abilities* and *functions*. To explain a cognitive function, we need only specify a mechanism that can perform the function. The methods of cognitive science are well-suited for this sort of explanation, and so are well-suited to the easy problems of consciousness. By contrast, the hard problem is hard precisely because it is not a problem about the performance of functions. The problem persists even when the performance of all the relevant functions is explained. (Here "function" is not used in the narrow teleological sense of something that a system is designed to do, but in the broader sense of any causal role in the production of behavior that a system

might perform.)

To explain reportability, for instance, is just to explain how a system could perform the function of producing reports on internal states. To explain internal access, we need to explain how a system could be appropriately affected by its internal states and use information about those states in directing later processes. To explain integration and control, we need to explain how a system's central processes can bring information contents together and use them in the facilitation of various behaviors. These are all problems about the explanation of functions.

How do we explain the performance of a function? By specifying a *mechanism* that performs the function. Here, neurophysiological and cognitive modeling are perfect for the task. If we want a detailed low-level explanation, we can specify the neural mechanism that is responsible for the function. If we want a more abstract explanation, we can specify a mechanism in computational terms. Either way, a full and satisfying explanation will result. Once we have specified the neural or computational mechanism that performs the function of verbal report, for example, the bulk of our work in explaining reportability is over.

In a way, the point is trivial. It is a *conceptual* fact about these phenomena that their explanation only involves the explanation of various functions, as the phenomena are *functionally definable*. All it *means* for reportability to be instantiated in a system is that the system has the capacity for verbal reports of internal information. All it means for a system to be awake is for it to be appropriately receptive to information from the environment and for it to be able to use this information in directing behavior in an appropriate way. To see that this sort of thing is a conceptual fact, note that someone who says "you have explained the performance of the verbal report function, but you have not explained reportability" is making a trivial conceptual mistake about reportability. All it could *possibly* take to explain reportability is an explanation of how the relevant function is performed; the same goes for the other phenomena in question.

Throughout the higher-level sciences, reductive explanation works in just this way. To explain the gene, for instance, we needed to specify the mechanism that stores and transmits hereditary information from one generation to the next. It turns out that DNA performs this function; once we explain how the function is performed, we have explained the gene. To explain life, we ultimately need to explain how a system can reproduce, adapt to its environment, metabolize, and so on. All of these are questions about the performance of functions, and so are well-suited to reductive explanation. The same holds for most problems in cognitive science. To explain learning, we need to explain the way in which a system's behavioral capacities are modified in light of environmental information, and the way in which new information can be brought to bear in adapting a system's actions to its environment. If we show how a neural or computational mechanism does the job, we have explained learning. We can say the same for other cognitive phenomena, such as perception, memory, and language. Sometimes the relevant functions need to be characterized quite subtly, but it is clear that insofar as cognitive science explains these phenomena at all, it does so by explaining the performance of functions.

When it comes to conscious experience, this sort of explanation fails. What makes the hard problem hard

and almost unique is that it goes *beyond* problems about the performance of functions. To see this, note that even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience - perceptual discrimination, categorization, internal access, verbal report - there may still remain a further unanswered question: *Why is the performance of these functions accompanied by experience?* A simple explanation of the functions leaves this question open.

There is no analogous further question in the explanation of genes, or of life, or of learning. If someone says "I can see that you have explained how DNA stores and transmits hereditary information from one generation to the next, but you have not explained how it is a *gene*", then they are making a conceptual mistake. All it means to be a gene is to be an entity that performs the relevant storage and transmission function. But if someone says "I can see that you have explained how information is discriminated, integrated, and reported, but you have not explained how it is *experienced*", they are not making a conceptual mistake. This is a nontrivial further question.

This further question is the key question in the problem of consciousness. Why doesn't all this information-processing go on "in the dark", free of any inner feel? Why is it that when electromagnetic waveforms impinge on a retina and are discriminated and categorized by a visual system, this discrimination and categorization is experienced as a sensation of vivid red? We know that conscious experience *does* arise when these functions are performed, but the very fact that it arises is the central mystery. There is an *explanatory gap* (a term due to Levine 1983) between the functions and experience, and we need an explanatory bridge to cross it. A mere account of the functions stays on one side of the gap, so the materials for the bridge must be found elsewhere.

This is not to say that experience *has* no function. Perhaps it will turn out to play an important cognitive role. But for any role it might play, there will be more to the explanation of experience than a simple explanation of the function. Perhaps it will even turn out that in the course of explaining a function, we will be led to the key insight that allows an explanation of experience. If this happens, though, the discovery will be an *extra* explanatory reward. There is no cognitive function such that we can say in advance that explanation of that function will *automatically* explain experience.

To explain experience, we need a new approach. The usual explanatory methods of cognitive science and neuroscience do not suffice. These methods have been developed precisely to explain the performance of cognitive functions, and they do a good job of it. But as these methods stand, they are *only* equipped to explain the performance of functions. When it comes to the hard problem, the standard approach has nothing to say.

4 Some case-studies

In the last few years, a number of works have addressed the problems of consciousness within the framework of cognitive science and neuroscience. This might suggest that the analysis above is faulty, but in fact a close examination of the relevant work only lends the analysis further support. When we investigate just which aspects of consciousness these studies are aimed at, and which aspects they end up

explaining, we find that the ultimate target of explanation is always one of the easy problems. I will illustrate this with two representative examples.

The first is the "neurobiological theory of consciousness" outlined by Crick and Koch (1990; see also Crick 1994). This theory centers on certain 35-75 hertz neural oscillations in the cerebral cortex; Crick and Koch hypothesize that these oscillations are the basis of consciousness. This is partly because the oscillations seem to be correlated with awareness in a number of different modalities - within the visual and olfactory systems, for example - and also because they suggest a mechanism by which the *binding* of information contents might be achieved. Binding is the process whereby separately represented pieces of information about a single entity are brought together to be used by later processing, as when information about the color and shape of a perceived object is integrated from separate visual pathways. Following others (e.g., Eckhorn *et al* 1988), Crick and Koch hypothesize that binding may be achieved by the synchronized oscillations of neuronal groups representing the relevant contents. When two pieces of information are to be bound together, the relevant neural groups will oscillate with the same frequency and phase.

The details of how this binding might be achieved are still poorly understood, but suppose that they can be worked out. What might the resulting theory explain? Clearly it might explain the binding of information contents, and perhaps it might yield a more general account of the integration of information in the brain. Crick and Koch also suggest that these oscillations activate the mechanisms of working memory, so that there may be an account of this and perhaps other forms of memory in the distance. The theory might eventually lead to a general account of how perceived information is bound and stored in memory, for use by later processing.

Such a theory would be valuable, but it would tell us nothing about why the relevant contents are experienced. Crick and Koch suggest that these oscillations are the neural *correlates* of experience. This claim is arguable - does not binding also take place in the processing of unconscious information? - but even if it is accepted, the *explanatory* question remains: Why do the oscillations give rise to experience? The only basis for an explanatory connection is the role they play in binding and storage, but the question of why binding and storage should themselves be accompanied by experience is never addressed. If we do not know why binding and storage should give rise to experience, telling a story about the oscillations cannot help us. Conversely, if we *knew* why binding and storage gave rise to experience, the neurophysiological details would be just the icing on the cake. Crick and Koch's theory gains its purchase by *assuming* a connection between binding and experience, and so can do nothing to explain that link.

I do not think that Crick and Koch are ultimately claiming to address the hard problem, although some have interpreted them otherwise. A published interview with Koch gives a clear statement of the limitations on the theory's ambitions.

Well, let's first forget about the really difficult aspects, like subjective feelings, for they may not have a scientific solution. The subjective state of play, of pain, of pleasure, of seeing blue, of smelling a rose - there seems to be a huge jump between the materialistic level, of explaining molecules and neurons, and the subjective level. Let's focus on things

that are easier to study - like visual awareness. You're now talking to me, but you're not looking at me, you're looking at the cappuccino, and so you are aware of it. You can say, 'It's a cup and there's some liquid in it.' If I give it to you, you'll move your arm and you'll take it - you'll respond in a meaningful manner. That's what I call awareness." ("What is Consciousness", *Discover*, November 1992, p. 96.)

The second example is an approach at the level of cognitive psychology. This is Baars' global workspace theory of consciousness, presented in his book *A Cognitive Theory of Consciousness*. According to this theory, the contents of consciousness are contained in a *global workspace*, a central processor used to mediate communication between a host of specialized nonconscious processors. When these specialized processors need to broadcast information to the rest of the system, they do so by sending this information to the workspace, which acts as a kind of communal blackboard for the rest of the system, accessible to all the other processors.

Baars uses this model to address many aspects of human cognition, and to explain a number of contrasts between conscious and unconscious cognitive functioning. Ultimately, however, it is a theory of *cognitive accessibility*, explaining how it is that certain information contents are widely accessible within a system, as well as a theory of informational integration and reportability. The theory shows promise as a theory of awareness, the functional correlate of conscious experience, but an explanation of experience itself is not on offer.

One might suppose that according to this theory, the contents of experience are precisely the contents of the workspace. But even if this is so, nothing internal to the theory *explains* why the information within the global workspace is experienced. The best the theory can do is to say that the information is experienced because it is *globally accessible*. But now the question arises in a different form: why should global accessibility give rise to conscious experience? As always, this bridging question is unanswered.

Almost all work taking a cognitive or neuroscientific approach to consciousness in recent years could be subjected to a similar critique. The "Neural Darwinism" model of Edelman (1989), for instance, addresses questions about perceptual awareness and the self-concept, but says nothing about why there should also be experience. The "multiple drafts" model of Dennett (1991) is largely directed at explaining the reportability of certain mental contents. The "intermediate level" theory of Jackendoff (1988) provides an account of some computational processes that underlie consciousness, but Jackendoff stresses that the question of how these "project" into conscious experience remains mysterious.

Researchers using these methods are often inexplicit about their attitudes to the problem of conscious experience, although sometimes they take a clear stand. Even among those who are clear about it, attitudes differ widely. In placing this sort of work with respect to the problem of experience, a number of different strategies are available. It would be useful if these strategic choices were more often made explicit.

The first strategy is simply to *explain something else*. Some researchers are explicit that the problem of

experience is too difficult for now, and perhaps even outside the domain of science altogether. These researchers instead choose to address one of the more tractable problems such as reportability or the self-concept. Although I have called these problems the "easy" problems, they are among the most interesting unsolved problems in cognitive science, so this work is certainly worthwhile. The worst that can be said of this choice is that in the context of research on consciousness it is relatively unambitious, and the work can sometimes be misinterpreted.

The second choice is to take a harder line and *deny the phenomenon*. (Variations on this approach are taken by Allport 1988, Dennett 1991, and Wilkes 1988.) According to this line, once we have explained the functions such as accessibility, reportability, and the like, there is no further phenomenon called "experience" to explain. Some explicitly deny the phenomenon, holding for example that what is not externally verifiable cannot be real. Others achieve the same effect by allowing that experience exists, but only if we equate "experience" with something like the capacity to discriminate and report. These approaches lead to a simpler theory, but are ultimately unsatisfactory. Experience is the most central and manifest aspect of our mental lives, and indeed is perhaps the key explanandum in the science of the mind. Because of this status as an explanandum, experience cannot be discarded like the vital spirit when a new theory comes along. Rather, it is the central fact that any theory of consciousness must explain. A theory that denies the phenomenon "solves" the problem by ducking the question.

In a third option, some researchers *claim to be explaining experience* in the full sense. These researchers (unlike those above) wish to take experience very seriously; they lay out their functional model or theory, and claim that it explains the full subjective quality of experience (e.g. Flohr 1992, Humphrey 1992). The relevant step in the explanation is usually passed over quickly, however, and usually ends up looking something like magic. After some details about information processing are given, experience suddenly enters the picture, but it is left obscure *how* these processes should suddenly give rise to experience. Perhaps it is simply taken for granted that it does, but then we have an incomplete explanation and a version of the fifth strategy below.

A fourth, more promising approach appeals to these methods to *explain the structure of experience*. For example, it is arguable that an account of the discriminations made by the visual system can account for the structural relations between different color experiences, as well as for the geometric structure of the visual field (see e.g., Clark 1992 and Hardin 1992). In general, certain facts about structures found in processing will correspond to and arguably explain facts about the structure of experience. This strategy is plausible but limited. At best, it takes the existence of experience for granted and accounts for some facts about its structure, providing a sort of nonreductive explanation of the structural aspects of experience (I will say more on this later). This is useful for many purposes, but it tells us nothing about why there should be experience in the first place.

A fifth and reasonable strategy is to *isolate the substrate of experience*. After all, almost everyone allows that experience *arises* one way or another from brain processes, and it makes sense to identify the sort of process from which it arises. Crick and Koch put their work forward as isolating the neural correlate of consciousness, for example, and Edelman (1989) and Jackendoff (1988) make related claims. Justification of these claims requires a careful theoretical analysis, especially as experience is not directly

observable in experimental contexts, but when applied judiciously this strategy can shed indirect light on the problem of experience. Nevertheless, the strategy is clearly incomplete. For a satisfactory theory, we need to know more than *which* processes give rise to experience; we need an account of why and how. A full theory of consciousness must build an explanatory bridge.

5 The extra ingredient

We have seen that there are systematic reasons why the usual methods of cognitive science and neuroscience fail to account for conscious experience. These are simply the wrong sort of methods: nothing that they give to us can yield an explanation. To account for conscious experience, we need an *extra ingredient* in the explanation. This makes for a challenge to those who are serious about the hard problem of consciousness: What is your extra ingredient, and why should *that* account for conscious experience?

There is no shortage of extra ingredients to be had. Some propose an injection of chaos and nonlinear dynamics. Some think that the key lies in nonalgorithmic processing. Some appeal to future discoveries in neurophysiology. Some suppose that the key to the mystery will lie at the level of quantum mechanics. It is easy to see why all these suggestions are put forward. None of the old methods work, so the solution must lie with *something* new. Unfortunately, these suggestions all suffer from the same old problems.

Nonalgorithmic processing, for example, is put forward by Penrose (1989; 1994) because of the role it might play in the process of conscious mathematical insight. The arguments about mathematics are controversial, but even if they succeed and an account of nonalgorithmic processing in the human brain is given, it will still only be an account of the *functions* involved in mathematical reasoning and the like. For a nonalgorithmic process as much as an algorithmic process, the question is left unanswered: why should this process give rise to experience? In answering *this* question, there is no special role for nonalgorithmic processing.

The same goes for nonlinear and chaotic dynamics. These might provide a novel account of the dynamics of cognitive functioning, quite different from that given by standard methods in cognitive science. But from dynamics, one only gets more dynamics. The question about experience here is as mysterious as ever. The point is even clearer for new discoveries in neurophysiology. These new discoveries may help us make significant progress in understanding brain function, but for any neural process we isolate, the same question will always arise. It is difficult to imagine what a proponent of new neurophysiology expects to happen, over and above the explanation of further cognitive functions. It is not as if we will suddenly discover a phenomenal glow inside a neuron!

Perhaps the most popular "extra ingredient" of all is quantum mechanics (e.g. Hameroff 1994). The attractiveness of quantum theories of consciousness may stem from a Law of Minimization of Mystery: consciousness is mysterious and quantum mechanics is mysterious, so maybe the two mysteries have a common source. Nevertheless, quantum theories of consciousness suffer from the same difficulties as neural or computational theories. Quantum phenomena have some remarkable functional properties, such

as nondeterminism and nonlocality. It is natural to speculate that these properties may play some role in the explanation of cognitive functions, such as random choice and the integration of information, and this hypothesis cannot be ruled out *a priori*. But when it comes to the explanation of experience, quantum processes are in the same boat as any other. The question of why these processes should give rise to experience is entirely unanswered.

(One special attraction of quantum theories is the fact that on some interpretations of quantum mechanics, consciousness plays an active role in "collapsing" the quantum wave function. Such interpretations are controversial, but in any case they offer no hope of *explaining* consciousness in terms of quantum processes. Rather, these theories *assume* the existence of consciousness, and use it in the explanation of quantum processes. At best, these theories tell us something about a physical role that consciousness may play. They tell us nothing about how it arises.)

At the end of the day, the same criticism applies to *any* purely physical account of consciousness. For any physical process we specify there will be an unanswered question: Why should this process give rise to experience? Given any such process, it is conceptually coherent that it could be instantiated in the absence of experience. It follows that no mere account of the physical process will tell us why experience arises. The emergence of experience goes beyond what can be derived from physical theory.

Purely physical explanation is well-suited to the explanation of physical *structures*, explaining macroscopic structures in terms of detailed microstructural constituents; and it provides a satisfying explanation of the performance of *functions*, accounting for these functions in terms of the physical mechanisms that perform them. This is because a physical account can *entail* the facts about structures and functions: once the internal details of the physical account are given, the structural and functional properties fall out as an automatic consequence. But the structure and dynamics of physical processes yield only more structure and dynamics, so structures and functions are all we can expect these processes to explain. The facts about experience cannot be an automatic consequence of any physical account, as it is conceptually coherent that any given process could exist without experience. Experience may *arise* from the physical, but it is not *entailed* by the physical.

The moral of all this is that *you can't explain conscious experience on the cheap*. It is a remarkable fact that reductive methods - methods that explain a high-level phenomenon wholly in terms of more basic physical processes - work well in so many domains. In a sense, one *can* explain most biological and cognitive phenomena on the cheap, in that these phenomena are seen as automatic consequences of more fundamental processes. It would be wonderful if reductive methods could explain experience, too; I hoped for a long time that they might. Unfortunately, there are systematic reasons why these methods must fail. Reductive methods are successful in most domains because what needs explaining in those domains are structures and functions, and these are the kind of thing that a physical account can entail. When it comes to a problem over and above the explanation of structures and functions, these methods are impotent.

This might seem reminiscent of the vitalist claim that no physical account could explain life, but the cases are disanalogous. What drove vitalist skepticism was doubt about whether physical mechanisms

could perform the many remarkable functions associated with life, such as complex adaptive behavior and reproduction. The conceptual claim that explanation of functions is what is needed was implicitly accepted, but lacking detailed knowledge of biochemical mechanisms, vitalists doubted whether any physical process could do the job and put forward the hypothesis of the vital spirit as an alternative explanation. Once it turned out that physical processes could perform the relevant functions, vitalist doubts melted away.

With experience, on the other hand, physical explanation of the functions is not in question. The key is instead the *conceptual* point that the explanation of functions does not suffice for the explanation of experience. This basic conceptual point is not something that further neuroscientific investigation will affect. In a similar way, experience is disanalogous to the *élan vital*. The vital spirit was put forward as an explanatory posit, in order to explain the relevant functions, and could therefore be discarded when those functions were explained without it. Experience is not an explanatory posit but an explanandum in its own right, and so is not a candidate for this sort of elimination.

It is tempting to note that all sorts of puzzling phenomena have eventually turned out to be explainable in physical terms. But each of these were problems about the observable behavior of physical objects, coming down to problems in the explanation of structures and functions. Because of this, these phenomena have always been the kind of thing that a physical account *might* explain, even if at some points there have been good reasons to suspect that no such explanation would be forthcoming. The tempting induction from these cases fails in the case of consciousness, which is not a problem about physical structures and functions. The problem of consciousness is puzzling in an entirely different way. An analysis of the problem shows us that conscious experience is just not the kind of thing that a wholly reductive account could succeed in explaining.

6 Nonreductive explanation

At this point some are tempted to give up, holding that we will never have a theory of conscious experience. McGinn (1989), for example, argues that the problem is too hard for our limited minds; we are "cognitively closed" with respect to the phenomenon. Others have argued that conscious experience lies outside the domain of scientific theory altogether.

I think this pessimism is premature. This is not the place to give up; it is the place where things get interesting. When simple methods of explanation are ruled out, we need to investigate the alternatives. Given that reductive explanation fails, *nonreductive* explanation is the natural choice.

Although a remarkable number of phenomena have turned out to be explicable wholly in terms of entities simpler than themselves, this is not universal. In physics, it occasionally happens that an entity has to be taken as *fundamental*. Fundamental entities are not explained in terms of anything simpler. Instead, one takes them as basic, and gives a theory of how they relate to everything else in the world. For example, in the nineteenth century it turned out that electromagnetic processes could not be explained in terms of the wholly mechanical processes that previous physical theories appealed to, so Maxwell and others

introduced electromagnetic charge and electromagnetic forces as new fundamental components of a physical theory. To explain electromagnetism, the ontology of physics had to be expanded. New basic properties and basic laws were needed to give a satisfactory account of the phenomena.

Other features that physical theory takes as fundamental include mass and space-time. No attempt is made to explain these features in terms of anything simpler. But this does not rule out the possibility of a theory of mass or of space-time. There is an intricate theory of how these features interrelate, and of the basic laws they enter into. These basic principles are used to explain many familiar phenomena concerning mass, space, and time at a higher level.

I suggest that a theory of consciousness should take experience as fundamental. We know that a theory of consciousness requires the addition of *something* fundamental to our ontology, as everything in physical theory is compatible with the absence of consciousness. We might add some entirely new nonphysical feature, from which experience can be derived, but it is hard to see what such a feature would be like. More likely, we will take experience itself as a fundamental feature of the world, alongside mass, charge, and space-time. If we take experience as fundamental, then we can go about the business of constructing a theory of experience.

Where there is a fundamental property, there are fundamental laws. A nonreductive theory of experience will add new principles to the furniture of the basic laws of nature. These basic principles will ultimately carry the explanatory burden in a theory of consciousness. Just as we explain familiar high-level phenomena involving mass in terms of more basic principles involving mass and other entities, we might explain familiar phenomena involving experience in terms of more basic principles involving experience and other entities.

In particular, a nonreductive theory of experience will specify basic principles telling us how experience depends on physical features of the world. These *psychophysical* principles will not interfere with physical laws, as it seems that physical laws already form a closed system. Rather, they will be a supplement to a physical theory. A physical theory gives a theory of physical processes, and a psychophysical theory tells us how those processes give rise to experience. We know that experience depends on physical processes, but we also know that this dependence cannot be derived from physical laws alone. The new basic principles postulated by a nonreductive theory give us the extra ingredient that we need to build an explanatory bridge.

Of course, by taking experience as fundamental, there is a sense in which this approach does not tell us why there is experience in the first place. But this is the same for any fundamental theory. Nothing in physics tells us why there is matter in the first place, but we do not count this against theories of matter. Certain features of the world need to be taken as fundamental by any scientific theory. A theory of matter can still explain all sorts of facts about matter, by showing how they are consequences of the basic laws. The same goes for a theory of experience.

This position qualifies as a variety of dualism, as it postulates basic properties over and above the

properties invoked by physics. But it is an innocent version of dualism, entirely compatible with the scientific view of the world. Nothing in this approach contradicts anything in physical theory; we simply need to add further *bridging* principles to explain how experience arises from physical processes. There is nothing particularly spiritual or mystical about this theory - its overall shape is like that of a physical theory, with a few fundamental entities connected by fundamental laws. It expands the ontology slightly, to be sure, but Maxwell did the same thing. Indeed, the overall structure of this position is entirely naturalistic, allowing that ultimately the universe comes down to a network of basic entities obeying simple laws, and allowing that there may ultimately be a theory of consciousness cast in terms of such laws. If the position is to have a name, a good choice might be *naturalistic dualism*.

If this view is right, then in some ways a theory of consciousness will have more in common with a theory in physics than a theory in biology. Biological theories involve no principles that are fundamental in this way, so biological theory has a certain complexity and messiness to it; but theories in physics, insofar as they deal with fundamental principles, aspire to simplicity and elegance. The fundamental laws of nature are part of the basic furniture of the world, and physical theories are telling us that this basic furniture is remarkably simple. If a theory of consciousness also involves fundamental principles, then we should expect the same. The principles of simplicity, elegance, and even beauty that drive physicists' search for a fundamental theory will also apply to a theory of consciousness.

(A technical note: Some philosophers argue that even though there is a *conceptual* gap between physical processes and experience, there need be no metaphysical gap, so that experience might in a certain sense still be physical (e.g. Hill 1991; Levine 1983; Loar 1990). Usually this line of argument is supported by an appeal to the notion of *a posteriori* necessity (Kripke 1980). I think that this position rests on a misunderstanding of *a posteriori* necessity, however, or else requires an entirely new sort of necessity that we have no reason to believe in; see Chalmers 1996 (also Jackson 1994 and Lewis 1994) for details. In any case, this position still concedes an *explanatory* gap between physical processes and experience. For example, the principles connecting the physical and the experiential will not be derivable from the laws of physics, so such principles must be taken as *explanatorily* fundamental. So even on this sort of view, the explanatory structure of a theory of consciousness will be much as I have described.)

7 Outline of a theory of consciousness

It is not too soon to begin work on a theory. We are already in a position to understand certain key facts about the relationship between physical processes and experience, and about the regularities that connect them. Once reductive explanation is set aside, we can lay those facts on the table so that they can play their proper role as the initial pieces in a nonreductive theory of consciousness, and as constraints on the basic laws that constitute an ultimate theory.

There is an obvious problem that plagues the development of a theory of consciousness, and that is the paucity of objective data. Conscious experience is not directly observable in an experimental context, so we cannot generate data about the relationship between physical processes and experience at will. Nevertheless, we all have access to a rich source of data in our own case. Many important regularities

between experience and processing can be inferred from considerations about one's own experience. There are also good indirect sources of data from observable cases, as when one relies on the verbal report of a subject as an indication of experience. These methods have their limitations, but we have more than enough data to get a theory off the ground.

Philosophical analysis is also useful in getting value for money out of the data we have. This sort of analysis can yield a number of principles relating consciousness and cognition, thereby strongly constraining the shape of an ultimate theory. The method of thought-experimentation can also yield significant rewards, as we will see. Finally, the fact that we are searching for a *fundamental* theory means that we can appeal to such nonempirical constraints as simplicity, homogeneity, and the like in developing a theory. We must seek to systematize the information we have, to extend it as far as possible by careful analysis, and then make the inference to the simplest possible theory that explains the data while remaining a plausible candidate to be part of the fundamental furniture of the world.

Such theories will always retain an element of speculation that is not present in other scientific theories, because of the impossibility of conclusive intersubjective experimental tests. Still, we can certainly construct theories that are compatible with the data that we have, and evaluate them in comparison to each other. Even in the absence of intersubjective observation, there are numerous criteria available for the evaluation of such theories: simplicity, internal coherence, coherence with theories in other domains, the ability to reproduce the properties of experience that are familiar from our own case, and even an overall fit with the dictates of common sense. Perhaps there will be significant indeterminacies remaining even when all these constraints are applied, but we can at least develop plausible candidates. Only when candidate theories have been developed will we be able to evaluate them.

A nonreductive theory of consciousness will consist in a number of *psychophysical principles*, principles connecting the properties of physical processes to the properties of experience. We can think of these principles as encapsulating the way in which experience arises from the physical. Ultimately, these principles should tell us what sort of physical systems will have associated experiences, and for the systems that do, they should tell us what sort of physical properties are relevant to the emergence of experience, and just what sort of experience we should expect any given physical system to yield. This is a tall order, but there is no reason why we should not get started.

In what follows, I present my own candidates for the psychophysical principles that might go into a theory of consciousness. The first two of these are *nonbasic principles* - systematic connections between processing and experience at a relatively high level. These principles can play a significant role in developing and constraining a theory of consciousness, but they are not cast at a sufficiently fundamental level to qualify as truly basic laws. The final principle is my candidate for a *basic principle* that might form the cornerstone of a fundamental theory of consciousness. This final principle is particularly speculative, but it is the kind of speculation that is required if we are ever to have a satisfying theory of consciousness. I can present these principles only briefly here; I argue for them at much greater length in Chalmers (1996).

1. The principle of structural coherence. This is a principle of coherence between the *structure of consciousness* and the *structure of awareness*. Recall that "awareness" was used earlier to refer to the various functional phenomena that are associated with consciousness. I am now using it to refer to a somewhat more specific process in the cognitive underpinnings of experience. In particular, the contents of awareness are to be understood as those information contents that are accessible to central systems, and brought to bear in a widespread way in the control of behavior. Briefly put, we can think of awareness as *direct availability for global control*. To a first approximation, the contents of awareness are the contents that are directly accessible and potentially reportable, at least in a language-using system.

Awareness is a purely functional notion, but it is nevertheless intimately linked to conscious experience. In familiar cases, wherever we find consciousness, we find awareness. Wherever there is conscious experience, there is some corresponding information in the cognitive system that is available in the control of behavior, and available for verbal report. Conversely, it seems that whenever information is available for report and for global control, there is a corresponding conscious experience. Thus, there is a direct correspondence between consciousness and awareness.

The correspondence can be taken further. It is a central fact about experience that it has a complex structure. The visual field has a complex geometry, for instance. There are also relations of similarity and difference between experiences, and relations in such things as relative intensity. Every subject's experience can be at least partly characterized and decomposed in terms of these structural properties: similarity and difference relations, perceived location, relative intensity, geometric structure, and so on. It is also a central fact that to each of these structural features, there is a corresponding feature in the information-processing structure of awareness.

Take color sensations as an example. For every distinction between color experiences, there is a corresponding distinction in processing. The different phenomenal colors that we experience form a complex three-dimensional space, varying in hue, saturation, and intensity. The properties of this space can be recovered from information-processing considerations: examination of the visual systems shows that waveforms of light are discriminated and analyzed along three different axes, and it is this three-dimensional information that is relevant to later processing. The three-dimensional structure of phenomenal color space therefore corresponds directly to the three dimensional structure of visual awareness. This is precisely what we would expect. After all, every color distinction corresponds to some reportable information, and therefore to a distinction that is represented in the structure of processing.

In a more straightforward way, the geometric structure of the visual field is directly reflected in a structure that can be recovered from visual processing. Every geometric relation corresponds to something that can be reported and is therefore cognitively represented. If we were given only the story about information-processing in an agent's visual and cognitive system, we could not *directly* observe that agent's visual experiences, but we could nevertheless infer those experiences' structural properties.

In general, any information that is consciously experienced will also be cognitively represented. The fine-

grained structure of the visual field will correspond to some fine-grained structure in visual processing. The same goes for experiences in other modalities, and even for nonsensory experiences. Internal mental images have geometric properties that are represented in processing. Even emotions have structural properties, such as relative intensity, that correspond directly to a structural property of processing; where there is greater intensity, we find a greater effect on later processes. In general, precisely because the structural properties of experience are accessible and reportable, those properties will be directly represented in the structure of awareness.

It is this isomorphism between the structures of consciousness and awareness that constitutes the principle of structural coherence. This principle reflects the central fact that even though cognitive processes do not conceptually entail facts about conscious experience, consciousness and cognition do not float free of one another but cohere in an intimate way.

This principle has its limits. It allows us to recover structural properties of experience from information-processing properties, but not all properties of experience are structural properties. There are properties of experience, such as the intrinsic nature of a sensation of red, that cannot be fully captured in a structural description. The very intelligibility of inverted spectrum scenarios, where experiences of red and green are inverted but all structural properties remain the same, show that structural properties constrain experience without exhausting it. Nevertheless, the very fact that we feel compelled to leave structural properties unaltered when we imagine experiences inverted between functionally identical systems shows how central the principle of structural coherence is to our conception of our mental lives. It is not a *logically* necessary principle, as after all we can imagine all the information processing occurring without any experience at all, but it is nevertheless a strong and familiar constraint on the psychophysical connection.

The principle of structural coherence allows for a very useful kind of indirect explanation of experience in terms of physical processes. For example, we can use facts about neural processing of visual information to indirectly explain the structure of color space. The facts about neural processing can entail and explain the structure of awareness; if we take the coherence principle for granted, the structure of experience will also be explained. Empirical investigation might even lead us to better understand the structure of awareness within a bat, shedding indirect light on Nagel's vexing question of what it is like to be a bat. This principle provides a natural interpretation of much existing work on the explanation of consciousness (e.g. Clark 1992 and Hardin 1992 on colors, and Akins 1993 on bats), although it is often appealed to inexplicitly. It is so familiar that it is taken for granted by almost everybody, and is a central plank in the cognitive explanation of consciousness.

The coherence between consciousness and awareness also allows a natural interpretation of work in neuroscience directed at isolating the *substrate* (or the *neural correlate*) of consciousness. Various specific hypotheses have been put forward. For example, Crick and Koch (1990) suggest that 40-Hz oscillations may be the neural correlate of consciousness, whereas Libet (1993) suggests that temporally-extended neural activity is central. If we accept the principle of coherence, the most *direct* physical correlate of consciousness is awareness: the process whereby information is made directly available for global control. The different specific hypotheses can be interpreted as empirical suggestions about how

awareness might be achieved. For example, Crick and Koch suggest that 40-Hz oscillations are the gateway by which information is integrated into working memory and thereby made available to later processes. Similarly, it is natural to suppose that Libet's temporally extended activity is relevant precisely because only that sort of activity achieves global availability. The same applies to other suggested correlates such as the "global workspace" of Baars (1988), the "high-quality representations" of Farah (1994), and the "selector inputs to action systems" of Shallice (1972). All these can be seen as hypotheses about the *mechanisms of awareness*: the mechanisms that perform the function of making information directly available for global control.

Given the coherence between consciousness and awareness, it follows that a mechanism of awareness will itself be a correlate of conscious experience. The question of just *which* mechanisms in the brain govern global availability is an empirical one; perhaps there are many such mechanisms. But if we accept the coherence principle, we have reason to believe that the processes that *explain* awareness will at the same time be part of the *basis* of consciousness.

2. The principle of organizational invariance. This principle states that any two systems with the same fine-grained *functional organization* will have qualitatively identical experiences. If the causal patterns of neural organization were duplicated in silicon, for example, with a silicon chip for every neuron and the same patterns of interaction, then the same experiences would arise. According to this principle, what matters for the emergence of experience is not the specific physical makeup of a system, but the abstract pattern of causal interaction between its components. This principle is controversial, of course. Some (e.g. Searle 1980) have thought that consciousness is tied to a specific biology, so that a silicon isomorph of a human need not be conscious. I believe that the principle can be given significant support by the analysis of thought-experiments, however.

Very briefly: suppose (for the purposes of a *reductio ad absurdum*) that the principle is false, and that there could be two functionally isomorphic systems with different experiences. Perhaps only one of the systems is conscious, or perhaps both are conscious but they have different experiences. For the purposes of illustration, let us say that one system is made of neurons and the other of silicon, and that one experiences red where the other experiences blue. The two systems have the same organization, so we can imagine gradually transforming one into the other, perhaps replacing neurons one at a time by silicon chips with the same local function. We thus gain a spectrum of intermediate cases, each with the same organization, but with slightly different physical makeup and slightly different experiences. Along this spectrum, there must be two systems *A* and *B* between which we replace less than one tenth of the system, but whose experiences differ. These two systems are physically identical, except that a small neural circuit in *A* has been replaced by a silicon circuit in *B*.

The key step in the thought-experiment is to take the relevant neural circuit in *A*, and install alongside it a causally isomorphic silicon circuit, with a switch between the two. What happens when we flip the switch? By hypothesis, the system's conscious experiences will change; from red to blue, say, for the purposes of illustration. This follows from the fact that the system after the change is essentially a

version of *B*, whereas before the change it is just *A*.

But given the assumptions, there is no way for the system to *notice* the changes! Its causal organization stays constant, so that all of its functional states and behavioral dispositions stay fixed. As far as the system is concerned, nothing unusual has happened. There is no room for the thought, "Hmm! Something strange just happened!". In general, the structure of any such thought must be reflected in processing, but the structure of processing remains constant here. If there were to be such a thought it must float entirely free of the system and would be utterly impotent to affect later processing. (If it affected later processing, the systems would be functionally distinct, contrary to hypothesis). We might even flip the switch a number of times, so that experiences of red and blue dance back and forth before the system's "inner eye". According to hypothesis, the system can never notice these "dancing qualia".

This I take to be a *reductio* of the original assumption. It is a central fact about experience, very familiar from our own case, that whenever experiences change significantly and we are paying attention, we can notice the change; if this were not to be the case, we would be led to the skeptical possibility that our experiences are dancing before our eyes all the time. This hypothesis has the same status as the possibility that the world was created five minutes ago: perhaps it is logically coherent, but it is not plausible. Given the extremely plausible assumption that changes in experience correspond to changes in processing, we are led to the conclusion that the original hypothesis is impossible, and that any two functionally isomorphic systems must have the same sort of experiences. To put it in technical terms, the philosophical hypotheses of "absent qualia" and "inverted qualia", while logically possible, are empirically and nomologically impossible.

(Some may worry that a silicon isomorph of a neural system might be impossible for technical reasons. That question is open. The invariance principle says only that *if* an isomorph is possible, then it will have the same sort of conscious experience.)

There is more to be said here, but this gives the basic flavor. Once again, this thought experiment draws on familiar facts about the coherence between consciousness and cognitive processing to yield a strong conclusion about the relation between physical structure and experience. If the argument goes through, we know that the only physical properties directly relevant to the emergence of experience are *organizational* properties. This acts as a further strong constraint on a theory of consciousness.

3. The double-aspect theory of information. The two preceding principles have been *nonbasic* principles. They involve high-level notions such as "awareness" and "organization", and therefore lie at the wrong level to constitute the fundamental laws in a theory of consciousness. Nevertheless, they act as strong constraints. What is further needed are *basic* principles that fit these constraints and that might ultimately explain them.

The basic principle that I suggest centrally involves the notion of *information*. I understand information

in more or less the sense of Shannon (1948). Where there is information, there are *information states* embedded in an *information space*. An information space has a basic structure of *difference* relations between its elements, characterizing the ways in which different elements in a space are similar or different, possibly in complex ways. An information space is an abstract object, but following Shannon we can see information as *physically embodied* when there is a space of distinct physical states, the differences between which can be transmitted down some causal pathway. The states that are transmitted can be seen as themselves constituting an information space. To borrow a phrase from Bateson (1972), physical information is a *difference that makes a difference*.

The double-aspect principle stems from the observation that there is a direct isomorphism between certain physically embodied information spaces and certain *phenomenal* (or experiential) information spaces. From the same sort of observations that went into the principle of structural coherence, we can note that the differences between phenomenal states have a structure that corresponds directly to the differences embedded in physical processes; in particular, to those differences that make a difference down certain causal pathways implicated in global availability and control. That is, we can find the *same* abstract information space embedded in physical processing and in conscious experience.

This leads to a natural hypothesis: that information (or at least some information) has two basic aspects, a physical aspect and a phenomenal aspect. This has the status of a basic principle that might underlie and explain the emergence of experience from the physical. Experience arises by virtue of its status as one aspect of information, when the other aspect is found embodied in physical processing.

This principle is lent support by a number of considerations, which I can only outline briefly here. First, consideration of the sort of physical changes that correspond to changes in conscious experience suggests that such changes are always relevant by virtue of their role in constituting *informational changes* - differences within an abstract space of states that are divided up precisely according to their causal differences along certain causal pathways. Second, if the principle of organizational invariance is to hold, then we need to find some fundamental *organizational* property for experience to be linked to, and information is an organizational property *par excellence*. Third, this principle offers some hope of explaining the principle of structural coherence in terms of the structure present within information spaces. Fourth, analysis of the cognitive explanation of our *judgments* and *claims* about conscious experience - judgments that are functionally explainable but nevertheless deeply tied to experience itself - suggests that explanation centrally involves the information states embedded in cognitive processing. It follows that a theory based on information allows a deep coherence between the explanation of experience and the explanation of our judgments and claims about it.

Wheeler (1990) has suggested that information is fundamental to the physics of the universe. According to this "it from bit" doctrine, the laws of physics can be cast in terms of information, postulating different states that give rise to different effects without actually saying what those states *are*. It is only their position in an information space that counts. If so, then information is a natural candidate to also play a role in a fundamental theory of consciousness. We are led to a conception of the world on which information is truly fundamental, and on which it has two basic aspects, corresponding to the physical and the phenomenal features of the world.

Of course, the double-aspect principle is extremely speculative and is also underdetermined, leaving a number of key questions unanswered. An obvious question is whether *all* information has a phenomenal aspect. One possibility is that we need a further constraint on the fundamental theory, indicating just what *sort* of information has a phenomenal aspect. The other possibility is that there is no such constraint. If not, then experience is much more widespread than we might have believed, as information is everywhere. This is counterintuitive at first, but on reflection I think the position gains a certain plausibility and elegance. Where there is simple information processing, there is simple experience, and where there is complex information processing, there is complex experience. A mouse has a simpler information-processing structure than a human, and has correspondingly simpler experience; perhaps a thermostat, a maximally simple information processing structure, might have maximally simple experience? Indeed, if experience is truly a fundamental property, it would be surprising for it to arise only every now and then; most fundamental properties are more evenly spread. In any case, this is very much an open question, but I believe that the position is not as implausible as it is often thought to be.

Once a fundamental link between information and experience is on the table, the door is opened to some grander metaphysical speculation concerning the nature of the world. For example, it is often noted that physics characterizes its basic entities only *extrinsically*, in terms of their relations to other entities, which are themselves characterized extrinsically, and so on. The intrinsic nature of physical entities is left aside. Some argue that no such intrinsic properties exist, but then one is left with a world that is pure causal flux (a pure flow of information) with no properties for the causation to relate. If one allows that intrinsic properties exist, a natural speculation given the above is that the intrinsic properties of the physical - the properties that causation ultimately relates - are themselves phenomenal properties. We might say that phenomenal properties are the internal aspect of information. This could answer a concern about the causal relevance of experience - a natural worry, given a picture on which the physical domain is causally closed, and on which experience is supplementary to the physical. The informational view allows us to understand how experience might have a subtle kind of causal relevance in virtue of its status as the intrinsic nature of the physical. This metaphysical speculation is probably best ignored for the purposes of developing a scientific theory, but in addressing some philosophical issues it is quite suggestive.

8 Conclusion

The theory I have presented is speculative, but it is a candidate theory. I suspect that the principles of structural coherence and organizational invariance will be planks in any satisfactory theory of consciousness; the status of the double-aspect theory of information is less certain. Indeed, right now it is more of an idea than a theory. To have any hope of eventual explanatory success, it will have to be specified more fully and fleshed out into a more powerful form. Still, reflection on just what is plausible and implausible about it, on where it works and where it fails, can only lead to a better theory.

Most existing theories of consciousness either deny the phenomenon, explain something else, or elevate the problem to an eternal mystery. I hope to have shown that it is possible to make progress on the

problem even while taking it seriously. To make further progress, we will need further investigation, more refined theories, and more careful analysis. The hard problem is a hard problem, but there is no reason to believe that it will remain permanently unsolved.[*]

*[[The arguments in this paper are presented in greater depth in my book *The Conscious Mind* (Oxford University Press, 1996). Thanks to Francis Crick, Peggy DesAutels, Matthew Elton, Liane Gabora, Christof Koch, Paul Rhodes, Gregg Rosenberg, and Sharon Wahl for their comments.]]

Further Reading

The problems of consciousness have been widely discussed in the recent philosophical literature. For some conceptual clarification of the various problems of consciousness, see Block 1995, Nelkin 1993, and Tye 1995. Those who have stressed the difficulties of explaining experience in physical terms include Hodgson 1988, Jackson 1982, Levine 1983, Lockwood 1989, McGinn 1989, Nagel 1974, Seager 1991, Searle 1991, Strawson 1994, and Velmans 1991, among others. Those who take a reductive approach include Churchland 1995, Clark 1992, Dennett 1991, Dretske 1995, Kirk 1994, Rosenthal 1996, and Tye 1995. There have not been many attempts to build detailed nonreductive theories in the literature, but see Hodgson 1988 and Lockwood 1989 for some thoughts in that direction. Two excellent collections of recent articles on consciousness are Block, Flanagan, and Güzeldeire 1996 and Metzinger 1995.

References

Akins, K. 1993. What is it like to be boring and myopic? In (B. Dahlbom, ed.) *Dennett and his Critics*. Oxford: Blackwell.

Allport, A. 1988. What concept of consciousness? In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

Baars, B.J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Bateson, G. 1972. *Steps to an Ecology of Mind*. Chandler Publishing.

Block, N. 1995. On a confusion about the function of consciousness. *Behavioral and Brain Sciences*.

Block, N, Flanagan, O. & Güzeldeire, G, (eds.) 1996. *The Nature of Consciousness: Philosophical and Scientific Debates*. Cambridge, MA: MIT Press.

Chalmers, D.J. 1996. *The Conscious Mind*. New York: Oxford University Press.

Churchland, P.M. 1995. *The Engine of Reason, The Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, MA: MIT Press.

Clark, A. 1992. *Sensory Qualities*. Oxford: Oxford University Press.

- Crick, F. and Koch, C. 1990. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2:263-275.
- Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Scribners.
- Dennett, D.C. 1991. *Consciousness Explained*. Boston: Little, Brown.
- Dretske, F.I. 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press.
- Edelman, G. 1989. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.
- Farah, M.J. 1994. Visual perception and visual awareness after brain damage: A tutorial overview. In (C. Umiltà and M. Moscovitch, eds.) *Consciousness and Unconscious Information Processing: Attention and Performance* 15. Cambridge, MA: MIT Press.
- Flohr, H. 1992. Qualia and brain processes. In (A. Beckermann, H. Flohr, and J. Kim, eds.) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. Berlin: De Gruyter.
- Hameroff, S.R. 1994. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1:91-118.
- Hardin, C.L. 1992. Physiology, phenomenology, and Spinoza's true colors. In (A. Beckermann, H. Flohr, and J. Kim, eds.) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. Berlin: De Gruyter.
- Hill, C.S. 1991. *Sensations: A Defense of Type Materialism*. Cambridge: Cambridge University Press.
- Hodgson, D. 1988. *The Mind Matters: Consciousness and Choice in a Quantum World*. Oxford: Oxford University Press.
- Humphrey, N. 1992. *A History of the Mind*. New York: Simon and Schuster.
- Jackendoff, R. 1987. *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.
- Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32: 127-36.
- Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, and S. White, eds.) *Philosophy and the Cognitive Sciences*. Vienna: Holder-Pichler-Tempsky.
- Kirk, R. 1994. *Raw Feeling: A Philosophical Account of the Essence of Consciousness*. Oxford: Oxford

University Press.

Kripke, S. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.

Lewis, D. 1994. Reduction of mind. In (S. Guttenplan, ed.) *A Companion to the Philosophy of Mind*. Oxford: Blackwell.

Libet, B. 1993. The neural time factor in conscious and unconscious events. In (G.R. Block and J. Marsh, eds.) *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Chichester: John Wiley and Sons.

Loar, B. 1990. Phenomenal states. *Philosophical Perspectives* 4:81-108.

Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford: Blackwell.

McGinn, C. 1989. Can we solve the mind-body problem? *Mind* 98:349-66.

Metzinger, T. 1995. *Conscious Experience*. Paderborn: Schöningh.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 4:435-50.

Nelkin, N. 1993. What is consciousness? *Philosophy of Science* 60:419-34.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Penrose, R. 1989. *The Emperor's New Mind*. Oxford: Oxford University Press.

Penrose, R. 1994. *Shadows of the Mind*. Oxford: Oxford University Press.

Rosenthal, D.M. 1996. A theory of consciousness. In (N. Block, O. Flanagan, and G. Güzelde, eds.) *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Seager, W.E. 1991. *Metaphysics of Consciousness*. London: Routledge.

Searle, J.R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-57.

Searle, J.R. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

- Shallice, T. 1972. Dual functions of consciousness. *Psychological Review* 79:383-93.
- Shannon, C.E. 1948. A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379-423.
- Strawson, G. 1994. *Mental Reality*. Cambridge, MA: MIT Press.
- Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.
- Velmans, M. 1991. Is human information-processing conscious? *Behavioral and Brain Sciences* 14:651-69.
- Wheeler, J.A. 1990. Information, physics, quantum: The search for links. In (W. Zurek, ed.) *Complexity, Entropy, and the Physics of Information*. Redwood City, CA: Addison-Wesley.
- Wilkes, K.V. 1988. - , Yishi, Duh, Um and consciousness. In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

Reply to Mulhauser's Review of *The Conscious Mind*

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

Gregory Mulhauser's vigorous [review](#) of my book *The Conscious Mind* includes a number of misrepresentations and philosophical misunderstandings. I would not like these to be perpetuated, so I will make a few comments in reply. I will go through the review, attempting to find objections of substance and to answer them where they are found. Amidst the rhetoric and the impressive repertoire of adjectives, I think a couple of interesting points on minor issues can be discerned.

First, I should clarify the notion of "taking consciousness seriously", which serves as a premise in my work. Mulhauser characterizes this as the assumption that no cognitive theory of consciousness will suffice. The latter assumption would indeed beg some crucial questions, but it is not the assumption that I make. I make an assumption about the problem of consciousness, not about any solution. To quote (p. xii):

Throughout the book, I have assumed that consciousness exists, and that to redefine the problem as that of explaining how certain cognitive and behavioral functions are performed is unacceptable. This is what I mean by taking consciousness seriously.

That is, the premise is simply that there is a phenomenon to be explained, and that the problems of explaining such functions as discrimination, integration, self-monitoring, reportability, and so on do not exhaust all the problems in the vicinity. The deepest problem of consciousness, as I understand it, is not the problem of how all these functions are performed, but rather the problem of explaining how and why all this activity supports states of subjective experience.

This isn't to make any assumptions about the nature of the solution: plenty of people agree with the premise but still think that one way or another they can get a cognitive or materialist theory of consciousness to work. Of course, I do go on to *argue* that if the premise is granted, it turns out that such theories will always be incomplete.

Like many people (materialists and dualists alike), I find this premise obvious, although I can no more "prove" it than I can prove that I am conscious. At the very least, to deny this premise would require extraordinarily strong arguments, of a type that I have never seen. In my experience the large majority of people find it obvious; but certainly there are some that deny it, and arguments over whether the premise

is true or false rapidly descend into table-pounding. Wishing to avoid that dead end, I prefer to simply state the assumption up front. Mulhauser complains that the assumption is not "discharged", but of course that is the whole point of making it a premise. I do argue for it where I can, but there is no denying that such arguments - on either side - ultimately come down to a bedrock of intuition at some point. The result, as I also say up front, is that the minority of people who don't see a "hard problem" aren't going to find the book of more than intellectual interest.

The first point that looks like a substantial objection to my analysis is in Mulhauser's discussion of Chapter 2, where he appears to object to my notion of logical necessity and its relation to formal systems. As elsewhere, he does not say what his objection is, but simply says that he objects. Objections of this sort are difficult to answer.

His discussion of the central Chapter 3 mostly centers around the conceivability of zombies. I say explicitly in the text that this conceivability argument is more inconclusive than some of the others (due to the difficulty of assessing claims about conceivability), and that it does not bear the central burden of the argument, but commentators have often focused on it nevertheless. Mulhauser's description of my zombies as "a particularly strong variety: a physical, functional, and *psychological* duplicate", with its emphasis, makes the claim seem stronger than it is: "psychological" is being used here in a stipulative sense that adds nothing over and above "functional". What is relevant to the arguments is that a zombie be a physical and functional duplicate; that's all.

Mulhauser's objection to the argument is the standard "how do we know we can really conceive this". He doesn't address the real point, which is the challenge to isolate an inconsistency in the notion, and in particular to demonstrate the aspect of the concept of consciousness which could ground an conceptual entailment from the physical facts to the facts about consciousness. He neither comes out and embraces a functional definition of consciousness (equating what needs to be explained with reportability, integration, or some such complex function), which would require facing up to all the problems with such a definition, nor does he even hint at another way that such an entailment might go through, which would require facing up to the fact (stressed in the book) that structure and function only ever adds up to more structure and function. So we are left in a position of assertion without enlightenment.

Mulhauser's discussion of epistemic asymmetry makes an interesting point: a logical entailment from physical facts to facts about consciousness requires a grasp of the concept of consciousness, so the mere fact that someone without consciousness wouldn't derive facts about consciousness from the physical facts can be explained away by the observation that they lack the concept. This is a minor point - there is still a striking disanalogy here between our epistemic access to consciousness and other phenomena - but nevertheless I should have rephrased things in a way that brings that disanalogy out more directly. In any case, I also point out the epistemic asymmetry of consciousness in ways that aren't vulnerable to this objection, e.g. by giving the observer the concept of consciousness and noting that the facts about consciousness in others remain underdetermined.

His second interesting point is the idea that the explanatory gap may arise for reasons of formal

undecidability or computational intractability. There are obvious problems to be addressed: either consciousness is a structural/functional concept, in which case (a) it is most unclear why the facts about it should be any more undecidable or intractable than the facts about any other such concept (learning, reportability, or whatever), and (b) the view will presumably be vulnerable to arguments just like mine, establishing a gap between the structural/functional concept and consciousness while leaving the "intractable" physical details to one side; or it is not a structural/functional concept, in which case there are systematic reasons why it cannot be logically necessitated by the physical facts, intractability/undecidability or no. So I think even this sort of position is vulnerable to the sort of critique I provide. But in any case it would be very interesting to see such a view worked out in detail, and the project of doing so would be sure to be enlightening. I do address this view briefly on my pp. 138-40 - not in much depth, as the view had not been explicitly set out in the literature as far as I know - but there is no doubt that there is more to say about it, and it would be a very useful service to see the view worked out systematically.

Mulhauser's discussion of "explanatory exclusion" is off the point. His claim that my intuitions about consciousness are based in "explanatory exclusion" intuitions that apply equally to planets, dolphins, and so on ignores the many pages of argument earlier in the book analyzing the disanalogy in depth. Indeed, none of the arguments I give about consciousness would have a hope of getting off the ground when applied to planets, dolphins, and the like.

His discussion of the epistemological problems of consciousness adds nothing of substance. There are, of course, serious problems here - the first-person epistemology of consciousness poses some of the deepest questions in the field, and questions to which I have at best tentative and incomplete answers. But Mulhauser gives no reason to believe that these answers must be bad ones; he simply reiterates the problems, as I have laid them out. He is of course absolutely right that many of these problems can be avoided if we embrace a reductive functionalist view of consciousness (where all that needs to be explained is discrimination, integration, reportability, and other complex functions). In fact, if we embrace such a view, almost *all* the problems of consciousness are removed! Those who think this sort of "solution" provides any enlightenment are welcome to it.

The entire second half of Mulhauser's review is thrown deeply astray by his "discovery" of my "little secret" that I think functional awareness is logically necessary for consciousness: a "discovery" based solely on my statement "awareness is necessary for consciousness". But in context it is obvious that natural necessity is at issue - what is under discussion are the bridging laws connecting processing and consciousness - and the sentence at the end of the paragraph makes it clear that logical necessity couldn't possibly be what is meant. Indeed, this "discovery" is so much at odds with everything else in the book that its invocation suggests that the reviewer is holding a principle of charity far in abeyance.

Mulhauser also seems to think that my suggestions for psychophysical bridging principles are *a priori* deductions from the single fact that consciousness exists. This is false: they are inferences drawn from apparent regularities between facts about functioning and facts about experience, especially in the first-person case. I start by drawing connections between (1) functional facts about reportability, behavioral control, and the like, and (2) phenomenological observations about the structure and properties of

experience in the first-person case. These connections are certainly quite coarse (I make no claim to be a sophisticated phenomenologist) and I would be all for a project of refining them. But in all of them phenomenological data - observations about the structure of experience, for example, or its presence or absence in different sorts of cases - play a significant role. In every case the relevant data go well beyond "consciousness exists"; Mulhauser is quite right that no substantial inferences about the nature of psychophysical laws could be drawn from that datum alone.

(Some may object to using phenomenological data, but I would argue that without an appeal to such data, a theory of conscious experience as such cannot even get off the ground - it will be a third-person theory that never connects to the first-person. And indeed I would argue that everyone who draws conclusions about experience from processing data is implicitly relying on bridging principles that are partly based on first-person knowledge of phenomenology; I am just trying to bring them into the open.)

Mulhauser's discussion of the fading and dancing qualia arguments makes a number of mistakes. The arguments cannot come close to establishing that absent or inverted qualia are logically impossible, for reasons I address directly: first, nothing in the arguments suggests that fading and dancing qualia scenarios are incoherent, and second, they all take the empirical fact of the existence of consciousness as a premise. Mulhauser somehow slides from this second observation of mine to the reiterated "conclusion" that my psychophysical principles are all deduced a priori from this observation, but this is an elementary mistake - to say that an argument takes something as a premise is not to say that it takes it as its *only* premise! Mulhauser also makes a lot of play out of a tension between my arguments here and my supposed view that awareness is logically necessary for consciousness; but as noted above, this view is a product of his imagination.

For what it's worth, Mulhauser's suggestion that this supposed view implies that dancing qualia are logically impossible commits another philosophical mistake: someone could believe that awareness is logically necessary for consciousness but not logically sufficient, and in particular could believe that facts about awareness do not logically entail facts about the intrinsic character of specific qualia.

Mulhauser's vigorous discussion of the last three chapters of the book contains few criticisms of substance, on examination. He complains that I don't appeal to his favorite sort of information theory, namely Chaitin's. I think Chaitin's work is fascinating (certainly it is intrinsically much more interesting than Shannon's), but it is an entirely different creature and I didn't see its application here. If Mulhauser can find a way to exploit it in a theory of consciousness, more power to him. His further observations (e.g., that some sorts of information processing seem to happen without consciousness) simply go over ground already covered in the book, without addressing my discussion.

His complaints about my chapter on strong AI rest on my "failure to grasp the vacuity of naive functional theorizing", and on his appeal to philosophers such as Putnam who have pointed out this "vacuity". Like many, I am not deeply impressed by these arguments, although they have some interest. I have a lengthy analysis of the Putnam considerations in my "Does a Rock Implement Every Finite State Automaton?" (Synthese 108:309-33, 1996), and the general issues are discussed in "A Computational Foundation for

the Study of Cognition (of which a shorter version appeared in *Minds and Machines*, 1994). Mulhauser provides nothing more than an appeal to authority here, though, so there is little I can say in reply. And although Mulhauser suggests that a giant look-up table might be constructed to satisfy my CSA definition, I would be very interested to see such a CSA that is naturally possible (remember, it is natural possibility that is relevant here). As I discuss in my paper on Putnam's argument, the only such "false implementations" are probably beyond the bounds of natural possibility.

Mulhauser says nothing of substance in objecting to my discussion of incompleteness - once again, he simply spends a long paragraph asserting vigorously that he has objections. The same goes for his discussion of discreteness vs. continuity, in the next paragraph. Those who understand the issues will know that the Siegelmann/Sontag machine he refers to does not threaten anything I say: like every other "super-Turing" machine that has been proposed, it requires an infinite resource or infinitely precise initial conditions in order to compute a super-Turing number-theoretic function. With bounded precision, or even with unbounded precision but a fixed finite initial state, it does no better than a Turing machine. As with many other places in the review, the argument strategy seems to be "unargued appeal to irrelevant authority".

In his discussion of my analysis of the Everett interpretation of quantum mechanics, Mulhauser asserts quite correctly that (1) a similar analysis might be available to a materialist (assuming a materialist theory of consciousness could work at all), and (2) that these considerations do not provide knockdown support for the interpretation (as I say in the book, they simply remove one sort of objection, and thus give it indirect support; the interpretation still has serious problems, and indeed I am not sure whether I believe it). He apparently intends these remarks as criticisms, but it is hard to see why. The chapter is not intended as religious warfare either for this interpretation or for dualism; it is simply an exploration of the issues.

I will end with a few comments on the more unusual aspects of Mulhauser's review. Mulhauser is concerned to point out everywhere that an argument I use has appeared in the literature before. This seems redundant, as I acknowledge my multiple debts right up front, and I cite every debt as the book goes along (it's not for nothing that the book has a 400-entry bibliography). Apparently he is spurred on by "the widespread publicity...suggesting it would contain startling new arguments and thought experiments". He would not be the first person to review the media rather than the book, but it would be nice to have the book reviewed for what it is.

In reality, it's a book written by a mortal, and a book which tries to present the arguments - some old, some new - as carefully and as completely as possible. There's no question that anti-reductionist arguments have been around a long time, and negative arguments with the same general thrust as mine have been seen before, though they have usually been presented in a very piecemeal way. In the book, I try to really make the case properly: maybe I succeed and maybe I fail, but I make no apologies for retreading old ground at those points where I think it is the best way to do things.

I must respond to one particularly odd remark of Mulhauser's in this context. A characterization of my

fading and dancing qualia arguments - "both of which, contrary to popular tale, have appeared in one form or another elsewhere in the literature" - seems to suggest that I have borrowed someone else's arguments unacknowledged. In fact both arguments are accompanied by citations to every item I know in the literature to which they bear even a remote similarity. The fading qualia gradual-replacement scenario is of course an old folk tale in the literature, and in the paragraph where it is introduced I cite a number of papers that use it (although they use it in somewhat different ways), and I footnote some more. The dancing qualia argument, as far as I know, was original with me in my 1993 Ph.D. dissertation, although there do exist a couple of distantly related scenarios used to argue for quite different conclusions, and a more closely related version was formulated independently by Arnold Zuboff in a 1994 article. So as with much of the rest of the book, there are a number of elements which are familiar, and a number which are new. That's the way it usually is in philosophy.

Finally, let me express my bemusement at the religious tone that is so common in this sort of discussion. I don't know why discussions of consciousness provoke this tone, which is not nearly so pronounced in other areas of science and philosophy. I don't feel any particular emotional commitment to my conclusions myself (I may be a materialist in my heart, if not in my mind); I have come to the conclusion that they more or less have to be true, for systematic reasons, but if someone thinks they have a better option, that's fine. I tried to write a book setting out the lay of the land as I see it, laying out what seemed to me to be a promising approach, and openly acknowledging the various problems my approach faces along the way. If someone disagrees, then I'm interested to hear about it. It's a fascinating area of intellectual inquiry. But point-scoring and unargued rhetoric are uninteresting. Mulhauser makes a couple of points of substance along the way. I suggest that we are all best served if substance is what we stick to.

References

- Chaitin, G.J. 1990. *Information, Randomness, and Incompleteness* (second edition). World Scientific.
- Chalmers, D.J. 1994. [A computational foundation for the study of cognition](#). PNP Technical Report 94-03, Washington University. Shorter version published as "On implementing a computation". *Minds and Machines* 4:391-402.
- Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D.J. 1996. [Does a rock implement every finite-state automaton?](#) *Synthese* 108:309-33.
- Putnam, H. 1988. *Representation and Reality*. MIT Press.
- Siegelmann, H.T. & Sontag, E.D. 1994. Analog computation via neural networks. *Theoretical Computer Science* 131:331-60.

Zuboff, A. 1994. What is a mind? In (P. French, T. Uehling, & H. Wettstein, eds) *Philosophical Naturalism*. University of Notre Dame Press.

Response to Searle

A response to John Searle's review of my book in the *New York Review of Books*, March 6, 1997. This appears as a letter in *NYRB*, May 15, 1997. Searle [responded](#); my more detailed [response in turn](#) is also online.

In my book *The Conscious Mind*, I deny a number of claims that John Searle finds "obvious", and I make some claims that he finds "absurd". But if the mind/body problem has taught us anything, it is that nothing about consciousness is obvious, and that one person's obvious truth is another person's absurdity. So instead of throwing around this sort of language, it is best to examine the claims themselves and the arguments that I give for them, to see whether Searle says anything of substance that touches them.

The first is my claim that consciousness is a nonphysical feature of the world. I resisted this claim for a long time, before concluding that it is forced on one by a sound argument. The argument is complex, but the basic idea is simple: the physical structure of the world - the exact distribution of particles, fields, and forces in spacetime - is logically consistent with the absence of consciousness, so the presence of consciousness is a further fact about our world. Searle says this argument is "invalid": he suggests that the physical structure of the world is equally consistent with the addition of flying pigs, but that it does not follow that flying is nonphysical.

Here Searle makes two elementary mistakes. First, he gets the form of the argument wrong. To show that flying is nonphysical, we would need to show that the world's physical structure is consistent with the *absence* of flying. From the fact that one can *add* flying pigs to the world, nothing follows. Second, the scenario he describes is not consistent. A world with flying pigs would have a lot of extra matter hovering meters above the earth, for example, so it could not possibly have the same physical structure as ours. Putting these points together: the idea of a world physically identical to ours but without flying, or without pigs, or without rocks, is self-contradictory. But there is no contradiction in the idea of a world physically identical to ours without consciousness, as Searle himself admits.

The underlying point is that the position of pigs - and almost everything else about the world - is logically derivable from the world's physical structure, but the presence of consciousness is not. So to explain why and how brains support consciousness, an account of the brain alone is not enough; to bridge the gap, one needs to add independent "bridging" laws. One can resist this conclusion only by adopting a hard-line deflationism about consciousness. That path has its own problems, but in any case it is not open to Searle, who holds that consciousness is irreducible. Irreducibility has its consequences. Consistency requires that one face them directly.

The next issue is my nonreductive functionalism. This bridging law claims that systems with the same functional organization have the same sort of conscious experiences. My detailed argument for this claim

is not recognizable in the trivial argument that Searle presents as mine and rebuts. The basic idea, presented in Chapter 7 of the book but ignored by Searle, is that if the claim is false, then there can be massive changes in conscious experience which a subject can never notice. (Searle's own position is rebutted on p. 258.) He also points to patients with Guillain-Barre syndrome as a counterexample to my claim, but this once again gets the logic wrong. My claim concerns functionally identical beings, so it is irrelevant to point to people who function differently. I certainly do not claim that beings whose functioning differs from ours are unconscious.

The final issue is panpsychism: the claim that some degree of consciousness is associated with every system in the natural world. Here Searle misstates my view: he says that I am "explicitly committed" to this position, when I merely explore it and remain agnostic; and he says incorrectly that it is an implication of property dualism and nonreductive functionalism. One can quite consistently embrace those views and reject panpsychism, so the latter could not possibly function as a "reductio ad absurdum" of the former. I note also that the view which I describe as "strangely beautiful", and which Searle describes as "strangely self-indulgent", is a view I reject.

I do argue that panpsychism is not as unreasonable as is often supposed, and that there is no knockdown argument against it. Searle helps confirm the latter claim: while protesting "absurdity", his arguments against panpsychism have no substance. He declares that to be conscious, a system must have the right "causal powers", which turn out to be the powers to produce consciousness: true, but trivial and entirely unhelpful. And he says that simple systems (such as thermostats) do not have the "structure" required for consciousness; but this is precisely the claim at issue, and he provides no argument to support it (if we knew what sort of structure were required for consciousness, the mind-body problem would be half-solved). So we are left where we started. Panpsychism remains counterintuitive, but it cannot be ruled out at the start of inquiry.

In place of substantive arguments, Searle provides gut reactions: every time he disagrees with a view I discuss, he calls it "absurd". In the case of panpsychism (a view not endorsed by me), many might agree. In other cases, the word is devalued: it is not even surprising, for example, that mental terms such as "perception" are ambiguous between a process and a subjective experience; and given that a trillion interacting neurons can result in consciousness, there is no special absurdity in the idea that a trillion interacting silicon chips or humans might do the same. I do bite one bullet, in accepting that brain-based explanations of behavior can be given that do not invoke or imply consciousness (although this is not to say that consciousness is causally irrelevant). But Searle's own view on irreducibility would commit him to this view too, if he could only draw the implication.

Once we factor out mistakes, misrepresentations, and gut feelings, we are left with not much more than Searle's all-purpose critique: "the brain causes consciousness". Although this mantra (repeated at least ten times) is apparently intended as a source of great wisdom, it settles almost nothing that is at issue. It is entirely compatible with all of my views: we just need to distinguish cause from effect, and to note that it does not imply that *only* the brain causes consciousness. Indeed, Searle's claim is simply a statement of the problem, not a solution. If one accepts it, the real questions are: Why does the brain cause consciousness? In virtue of which of its properties? What are the relevant causal laws? Searle has nothing

to say about these questions. A real answer requires a theory: not just a theory of the brain, but also a detailed theory of the laws that bridge brain and consciousness. Without fulfilling this project, on which I make a start in my book, our understanding of consciousness will always remain at a primitive level.

David Chalmers
University of California, Santa Cruz

Go to:

- [David Chalmers' home page](#)
- [The Conscious Mind web page](#)
- [On "Consciousness and the Philosophers"](#) (second response to Searle)

On "Consciousness and the Philosophers"

[David J. Chalmers](#)

John Searle's [review](#) of my book *The Conscious Mind* appeared in the March 6, 1997 edition of the *New York Review of Books*. I replied in a [letter](#) printed in their May 15, 1997 edition, and Searle's [response](#) appeared simultaneously. I set up this web page so that interested people can see my reply to Searle in turn, and to give access to other relevant materials.

John Searle's reply to my letter continues the odd combination of mistakes, misrepresentations, and unargued gut reactions found in his original review. I will go through his response and address all his points, major and minor. I will also include a few comments on the original review which could not be included in the letter for reasons of space.

(1) Searle says

Chalmers thinks each of the psychological words "pain", "belief", etc., has two completely independent meanings, one where it refers to nonconscious functional processes and one where it refers to states of consciousness.

To clarify: I don't say that mental terms are ambiguous between consciousness and "nonconscious functional processes". I say that many such terms have two interpretations (not "completely independent" interpretations), one of which requires consciousness and one of which doesn't. For example, in the strong sense, "perception" has to be conscious; in the weak sense, it does not. So "subliminal perception" will count as "perception" in the weak but not the strong sense. I think this is just common sense. Note that the weak sense doesn't require that a process be "nonconscious"; rather, it is neutral on the question of consciousness. Subliminal perception and conscious perception count equally as perception in the weak sense; what matters isn't consciousness or its absence, but the way a state functions. So Searle is wrong to suggest that this distinction implies that there is a nonconscious "pain" correlated with every conscious "pain".

(2) On the question of explanatory irrelevance of consciousness to physical actions. It is true that I bite this bullet, in a certain sense. But the view can be made to seem stranger than it is. So I should make some clarifications.

First, I do not say that consciousness is *causally* irrelevant to action. That is a question I am neutral on, and I think there are interesting views that give consciousness causal relevance without doing any damage to the scientific world view. (For example, by making consciousness correspond to the intrinsic aspect of physical states which physics characterizes only extrinsically.)

Second, I do not say that consciousness can never be used to explain action, so that explanations that involve consciousness are invalid. I simply say that invoking consciousness is not *necessary* to explain actions; there will always be a physical explanation that does not invoke or imply consciousness. A better phrase would have been "explanatorily superfluous", rather than "explanatorily irrelevant". Something can be superfluous and still be relevant.

Third, it isn't true that my view implies that "if you think you are reading because you consciously want to read, you are mistaken". On my view, it is very likely that you are reading because you want to read. It is just that fact that the wanting is consciously experienced is not required for the explanation to go through. Conscious wants can explain actions, and nonconscious wants can explain actions too. Similarly, you drink because you are thirsty, but the consciously experienced aspect of that thirst needn't be appealed to to explain your action.

Of course there remains something counterintuitive about the explanatory superfluity of consciousness. But I think one is forced to it by sound arguments. And in the book I defend it at length, arguing that it is *merely* counterintuitive and has no fatal flaws. Searle does not address this discussion.

(3) On panpsychism (the view that consciousness is everywhere): In Searle's review he said I was "explicitly committed" to this view. I corrected this and said I was agnostic about the view, although I do explore it at length. Searle apparently thinks I am lying about this:

The most astonishing thing in Chalmers's letter is the claim that he did not "endorse" panpsychism, that he is "agnostic" about it. Well, in his book he presents extensive arguments for it and defenses of it. Here is what he actually says. First, he tells us that "consciousness arises from functional organization" (p. 249). And what is it about functional organization that does the job? It is, he tells us, "information", in the special sense of that word according to which everything in the world has information in it. "We might put this by suggesting as a basic principle that information (in the actual world) has two aspects, a physical and a phenomenal aspect" (p. 286). The closest he gets to agnosticism is this: "I do not have any knockdown arguments to prove that information is the key to the link between physical processes and conscious experience," but he immediately adds, "but there are some indirect ways of giving support to the idea." Whereupon he gives us several arguments for the double-aspect principle (p. 287). He hasn't proven panpsychism, but it is a hypothesis he thinks is well-supported. Since information in his sense is everywhere then consciousness is everywhere. If he argues that functional organization gives rise to consciousness, that it does so in virtue of information, that anything that has information would be conscious and that everything has information, then he is arguing that everything is conscious, by any logic that I am aware of.

Searle clearly does not grasp what is going on in this chapter of the book. First, with regard to the double-aspect theory of information: I am not committed even to this view, and indeed I say in the chapter (p. 310) that it is more likely wrong than right. But further, it simply isn't the case that the double-aspect view implies panpsychism. In the passage setting out the double-aspect view on p. 286, Searle has

apparently missed the sentence two sentences after the one he quotes:

Conversely, for *at least some* physically realized information spaces, whenever an information state in that space is realized physically, it is also realized phenomenally. [TCM, p. 286; italics added.]

And a couple of sentences later:

In fleshing the principle out into a theory, all sorts of questions need to be answered. For example, to just *which* physically realized information spaces does the basic principle apply? I will discuss this question further in section 4, but in the meantime I leave it open. [TCM, p. 286]

In section 4, I go on to say that there are two quite different ways in which the view can be developed. One can go for a *constrained* version of the view, putting constraints on the kind of information which has a conscious aspect; or one can go for an *unconstrained* version, on which all information has a conscious aspect. Only the latter implies panpsychism. In the book I explore both these options and remain neutral between them.

Searle asks

If he is not supporting panpsychism, why does he devote an entire section of a chapter to "What is it like to be a thermostat?", in which he describes the conscious life of thermostats?

Here he has simply lost the context. I spend some time exploring both the unconstrained and constrained views, considering how they might be developed and drawing out their implications. In the pages exploring the unconstrained view, I argue that panpsychism is not an unreasonable view, I defend it against various criticisms that might be raised, and I try to elaborate on consequences such as how we might conceive of the conscious experience of a thermostat, just as I elaborate on the constrained view after that. Apparently he mistakes this exploration of the consequences of one possible view for a statement of what I explicitly endorse. A quick glance at the context should rule out this interpretation: e.g., I say explicitly that I am considering both versions (p. 293), and I say explicitly that I consider the question of panpsychism open (p. 299).

I do say that I think there aren't any compelling arguments *against* panpsychism, and once again Searle helps to confirm this view. He tells us the only systems that we "know for a fact" are conscious are living systems with a certain sort of nervous systems. This is quite true, but it does not imply that simple systems are not conscious; it simply implies that we do not know for a fact that they are. So this might be grounds for agnosticism about panpsychism, but it is not grounds for rejection. He repeats his remarks about thermostats and such "not being remote candidates for consciousness", but this is again just assertion without argument. And he says that what goes on in a thermostat is quite different from the "specific features" in a brain: true, but to assume that these specific features are *required* for consciousness is again to beg the question.

(4) Searle apparently does not believe me when I say that the view that I characterize as "strangely beautiful" in the book is a view I rejected. He appears not to have read the relevant passages closely enough. The view so characterized is not, as Searle suggests, the view that "the universe consists of little bits of consciousness"; that is indeed a view I am agnostic about. Rather, it is the view on which the universe consists of "pure information" without any intrinsic nature (p. 303). This is a very different view, entirely incompatible with the other view, and I reject it on the grounds that some intrinsic nature is needed for the universe to be a universe at all.

(5) Searle's discussion of the argument for property dualism once again makes elementary mistakes. I argue that there is a logically possible world that is physically identical to ours without consciousness, and draw the conclusion that consciousness is a nonphysical feature of the world. Searle objects because this world would have different laws from ours.

Quite so, but in order to imagine such a world, you have to imagine a change in the laws of nature, a change in the laws by which physics and biology cause and realize consciousness. [...] From the facts of nature, including the laws, you can derive logically that this brain must be conscious.

Of course this is true, but it is also irrelevant. The point is that to derive the facts about consciousness, the *physical* facts and laws are not enough. You need to add in the *bridging* laws relating physical process and consciousness. Whereas in almost any other domains, you don't need any bridging laws: the physical facts and laws alone are enough to derive the position of pigs, the facts about digestion, the shape of rocks, the functioning of living processes, and so on. This is witnessed by the logical impossibility of a world physically identical to ours but in which rocks have different shape or in which pigs fly, and similarly for all the others. So Searle has simply evaded the point. Indeed, he has in effect conceded the point, by agreeing that a world physically identical to ours but without consciousness is logically possible. To be consistent, he too ought to be a property dualist.

Searle also suggests that I beg the question:

If consciousness is a physical feature of brains, then the absence of consciousness is *also* a change in the physical features of the world. That is, his argument works to establish property dualism only if it assumes that consciousness is not a physical feature, but that is what the argument was supposed to prove.

But nothing in the argument requires any such assumption. We need only stipulate that the world we are considering is a world identical in the facts and laws of microphysics, characterizing the exact distribution of particles, fields, forces, and so on, in space and time. We stipulate nothing one way or the other about consciousness, rocks, pigs, and the like. The point is that from these facts one can *derive* the facts about rocks, pigs, and the like but one can't derive the facts about consciousness, which is the relevant disanalogy. This argument *implies* that consciousness is not a physical feature of the world, but it does not *assume* it. (Except in the very weak sense in which the premises of *any* argument assume the conclusion: if the conclusion were false, one of the premises would have to be false!)

Of course I agree with Searle that *if* one builds in facts and laws about consciousness to the basic package of facts and laws from which one can derive everything else (facts about fields, particles, forces, etc), then one can derive the facts about consciousness. But that conclusion is trivial. Indeed, to include consciousness in this basic package is precisely to endorse the view that I hold. To avoid this conclusion, Searle has to derive the facts about consciousness from the set of physical facts and laws alone, and that is what he can't do.

(6) In discussing my arguments for nonreductive functionalism, Searle once again ignores the actual argumentation in the book. To recap: I argue that if consciousness could vary independently of functional organization, then there could be changes in consciousness (e.g. by replacing neurons by silicon) that a subject could never notice. The subject would certainly insist that "Nothing has changed", for example; this is a trivial consequence of the assumption that functional organization is preserved. Searle's position is that although the subject would not produce any "noticing behavior", they might nevertheless notice the change but be unable to express it in behavior. They might feel that they are paralyzed inside a body out of their control, for example.

Searle says that I merely "assume" this option is untenable, by assuming that there must be a match between conscious noticing and noticing behavior. But in fact I *argue* against this position at some length, around p. 258-59 of the book (as I noted in my letter). In particular, I argue that the view Searle wants to maintain requires a particularly bizarre and arbitrary law connecting physical states and belief contents. Searle does not address these arguments anywhere. Interested readers can find a version of my arguments for nonreductive functionalism in my online paper ["Absent Qualia, Fading Qualia, Dancing Qualia"](#).

(7) In Searle's original review, he uses patients with Guillain-Barre syndrome to argue against my nonreductive functionalism, by noting that their functional organization is "inappropriate" but that they are conscious anyway. I responded that this gets the logic wrong:

He also points to patients with Guillain-Barre syndrome as a counterexample to my claim, but this once again gets the logic wrong. My claim concerns functionally identical beings, so it is irrelevant to point to people who function differently. I certainly do not claim that beings whose functioning differs from ours are unconscious.

Searle replies:

He says incorrectly that patients with Guillain-Barre syndrome who have consciousness but the wrong functional organization are irrelevant, because they "function differently". But differently from what? As far as their physical behavior is concerned they function exactly like people who are totally unconscious, and thus on his definition they have exactly the same "functional organization" as unconscious people, even though they are perfectly conscious. Remember "functional organization" for him is always nonconscious. The Guillan-Barre patients have the

same functional organization, but different consciousness. Therefore there is no perfect match between functional organization and consciousness. Q.E.D.

Searle's point here is subtly different from his original version, but it is equally fallacious. Patients with Guillain-Barre syndrome certainly do not have the same functional organization as unconscious people. There is all sorts of complex functioning going on inside their heads that is not present in unconscious people. They have the same *behavior*, but that is all. Searle knows well that functional organization and behavior are quite different things. Indeed, the definition of functional organization that I give (p. 247) involves much more than mere behavior. And indeed the move from behaviorism to functionalism in the philosophy of mind was made in part for this reason: to allow more fine-grained distinctions than mere behavior could capture, and thus to handle cases such as those of perfect actors and paralytics. If Searle knows any elementary philosophy of mind, he knows this. But it renders his point entirely invalid.

(8) Searle finishes by saying that as candidates for explaining consciousness,

"functional organization" and "information" are nonstarters, because as he uses them, they have no causal explanatory power. To the extent you make the function and the information specific, they exist only relative to observers and interpreters.

But this claim is quite false. Searle has made it a number of times, generally without any substantive supporting argument. I argue in Chapter 9 of the book, and in more detail in my papers ["A Computational Foundation for the Study of Cognition"](#) and ["Does a Rock Implement Every Finite-State Automaton?"](#) that the relevant notions can be made perfectly precise with objective criteria, and are therefore not at all observer-relative. If a given system has a given functional organization, implements a given computation, and therefore realizes certain information, it does so as a matter of objective fact. Searle does not address these arguments at all.

Searle does make one point that might look like supporting argument:

If you strip away the observers and interpreters then the notions become empty, because now everything has information in it and has some sort of "functional organization" or other.

The latter part of this claim is true, but it no more implies that the notions are empty than the claim that everything has mass would imply that the notion of mass is empty. Everything has *some* organization, and carries *some* information, but the particular organizations and information realized by a given system will vary from system to system. A theory of the consciousness associated with a given system will appeal not just to the fact that *some* organization or information is present - that would be empty - but to the specific organizational and informational properties in a given case. (By analogy, an account of the sun's powerful gravitational force appeals not just to the fact that the sun has mass - that would be empty - but to the specific large mass that the sun has.) These specific properties will only be present in a small number of cases, so there is no danger of vacuity. This issue is discussed further in the [computation](#) paper mentioned above.

(9) Finally, are my property dualism and my functionalism compatible? As Searle defines the positions in the first paragraph, they are obviously inconsistent. He defines functionalism as the view that mental states consist in physical functional states, which obviously contradicts the view that mental states are not physical states. My own variety of "functionalism" is very different from this. I don't say that mental states *are* functional states, merely that they are supported by functional states in the sense that any two functionally identical beings (in the actual world) will have the same conscious states. To make an analogy: almost any property dualist will hold that consciousness is somehow supported by brain states, in the sense that two creatures with identical brains will have identical experiences, but this is not to say that consciousness *is* the brain state. Likewise, my own view is not that consciousness *is* the functional state. I argue against that view at length, e.g. on the grounds that it remains logically possible that a being could have the functional state without consciousness. But it is nevertheless *empirically* impossible to have the physical/functional state without consciousness, on my view; there is a psychophysical law connecting the two.

In his review, Searle suggested that there is a deep tension between my property dualism and my functionalism. Once the nonreductive nature of my functionalism is noted, this claim can be seen to be misguided. Property dualism is entirely neutral on the question of what *sort* of physical system supports consciousness, and on the precise nature of the laws that connect the two. Presumably the laws will say that consciousness arises from physical systems in virtue of certain of their properties. I simply argue that these properties are organizational properties, so that consciousness might arise equally from neurons and from silicon, for example. There is no tension here. (One might find more tension in a property dualist who held that consciousness is specifically biological!) If one takes the trouble to distinguish the separate issues at play here, it becomes clear that the two views are complementary, not contradictory.

Go to:

- [David Chalmers' home page](#)
- [The Conscious Mind web page](#)
- [Reviews of The Conscious Mind](#)
- [Papers on consciousness \(David Chalmers\)](#)

Online discussions of my work

Here are a few places where my work is discussed on the web. A separate page is devoted to [reviews of *The Conscious Mind*](#). For replies to some of these articles, see my page of [responses](#).

Papers on consciousness

- [A Materialist Response to David Chalmers' *The Conscious Mind*](#) (Paul Raymore)
- [Against Chalmers' Epiphenomenalism](#) (Glenn Braddock)
- [Are Zombies Positively Conceivable?](#) (Farid Masrour)
- [Chalmers on Consciousness and Quantum Mechanics](#) (Alex Byrne & Ned Hall) [[reply](#)]
- [Chalmers on the Justification of Phenomenal Judgment](#) (Tim Bayne) [[reply](#)]
- [Conceivability Arguments, or Revenge of the Zombies](#) (Katalin Balog) [[reply](#)]
- [Concepts and Consciousness](#) (Stephen Yablo) [[reply](#)]
- [Conceptual Analysis and the Explanatory Gap](#) (Ned Block & Robert Stalnaker) [[reply](#)]
- [Consciousness: Chalmers and Dennett](#) (Steve Torrance)
- [Consciousness and the Philosophers](#) (John Searle) [[reply 1](#) and [reply 2](#)]
- [Consciousness, Science, and the Nature of Explanation](#) (Marion Gothier)
- [Cosmic Consciousness for Tough Minds](#) (David Pearce)
- [Critique of the Conceivability Debate](#) (Douglas Alan)
- [Facing Backwards on the Problem of Consciousness](#) (Dan Dennett) [[reply](#)]
- [Function and Phenomenology: Closing the Explanatory Gap](#) (Tom Clark) [[reply](#)]
- [In the Hands of Zombies](#) (Navin Kartik)
- [Interdisciplinary Developments: Toward a Science of Consciousness](#) (P.R.F. Brown)
- [Letters to the Editor](#) (Scientific American)
- [Logical Possibility, Laws of Nature, and Mind](#) (Wallace Matson)
- [Mind and Experience](#) (Robert Pallbo)
- [Mind as Metaphor](#) (James Hopkins)
- [Reining in Chalmers: On the logical possibility of zombies](#) (Chris Mathieson)
- [Sniffing the Camembert: On the Conceivability of Zombies](#) (Allin Cottrell)
- [Textbook Kripkeanism and the Open Texture of Language](#) (Stephen Yablo).
- [The Fantasy of First-Person Science](#) (Daniel Dennett)
- [The Hard Problem is Dead: Long Live the Hard Problem](#) (Teed Rockwell)
- [The Hard Problem Isn't](#) (Michael Lissack)
- [The Hornswoggle Problem](#) (Pat Churchland) [[reply](#)]
- [The Relation of Consciousness to the Material World](#) (Max Velmans) [[reply](#)]

- [Zombie Killer](#) (Nigel Thomas)

Papers on mind, meaning, and modality:

- [A Conceivability Argument](#) (Adam Elga)
- [Bad Intensions](#) (Alex Byrne and Jim Pryor)
- [Chalmers on Epistemic Content](#)
- [Chalmers on the Apriority of Modal Knowledge](#) (Christopher Hill) [[reply](#)]
- [Cosmic Hermeneutics](#) (Alex Byrne) [[reply](#)]
- [Coulda, Woulda, Shoulda](#) (Stephen Yablo)
[[reply](#)]
- [Modal Rationalism and Logical Empiricism: Some Similarities](#) (Stephen Yablo)
[[reply](#)]
- [On Considering a Possible World as Actual](#) (Robert Stalnaker)
[[reply](#)]
- [Phenomenal Qualities and Phenomenal Concepts](#) (Norman Teng)
- [Possible Worlds Semantics for Belief Sentences](#) (Joe Lau)
- [What is A Priori and What is it Good For?](#) (David Henderson and Terry Horgan)
- [Why Positive and Negative Conceivability Can't Save the Conceivability-Possibility Link](#) (Matthew Phillips)

Papers on other subjects (AI, physics, extended mind):

- [A Subsymbolic Computationalist Reply to the Chinese Room](#) (Ken Marable)
- [Beyond the Doubting of a Shadow](#) (Roger Penrose)
- [Connectionism and the Language of Thought](#) (Murat Aydede)
- [Interpretations of Quantum Mechanics](#) (Jacques Mallah)
- [Learning Recursive Distributed Representations for Holistic Computation](#) (Lonnie Chrisman)
- [Structure-Mapping vs. High-Level Perception: The Mistaken Fight over the Explanation of Analogy](#) (Clayton Morrison and Eric Dietrich)
- ["The Extended Mind" -- Extended](#) (Joseph Fulda)
- [What Ain't in the Head](#) (Tim Maletic)

Other online discussion

- [Metaphysics of Consciousness: David Chalmers' *The Conscious Mind* in Historical and Contemporary Perspective](#). This was a conference at Buffalo in November 1999. The site includes a number of online papers.

- [Much Ado About Consciousness: An Interview with David Chalmers](#). This is an in-depth interview conducted by Andrew Chrucky (who has a very nice [philosophical website](#) of his own). This appeared in the Summer/Autumn 1998 issue of *Philosophy Now*, and goes into some interesting philosophical matters.
- An [interview](#) with me conducted by Stephen Jones at Tucson II, as part of his interesting [Brain Project](#). This site also has relevant discussions of the [hard problem](#) and of the [neural correlate of consciousness](#).
- [Mind your TOEs: An interview with David Chalmers](#). This interview was conducted by Trevor Thompson, an Australian science journalist, for 21C magazine. It's part of his [100% Sci-Phi](#) website.
- [A new form of materialism?](#) This is a novelty: a news article that goes into depth about serious analytic philosophy, at the website of [Philosophy News](#). The article is written by Barbara Montero, and discusses the symposium in *Philosophy and Phenomenological Research* on my book.
- An audio file of my [talk at Tucson II](#), entitled "On the Search for the Neural Correlate of Consciousness. (This is part of the extensive coverage of the conference by the [Times Higher Education Supplement](#), who also have most of the other talks.) You'll need a [RealAudio](#) player; note that the talk itself doesn't start until about the two-minute mark. To go along with the audio, here are some [visuals](#).
- An audio file of a [news piece on consciousness](#) broadcast on NPR's [All Things Considered](#) on September 17, 1996. Apart from me, it features Dan Dennett, Christof Koch, Irene Pepperberg and her parrot Alex. Once again, you'll need [RealAudio](#), and you also need to register.
- Another [audio piece](#), broadcast on [The Infinite Mind](#) in March 2001. I'm in a discussion with Christof Koch and Fred Goodwin. The program also features Descartes, Pat Churchland, and the Borg.
- Here is a [video piece](#) of a talk I gave at a [symposium on consciousness](#) at NIMH in May 2000. You can also find talks by Christof Koch, Allan Hobson, and Patricia Churchland, plus discussion.
- [Closer to Truth](#) is a sort of intellectual talk show broadcast on PBS. From their [Brain and Mind](#) page, you can find video movies and transcripts of three shows on which I appeared, plus other things on mind-brain-consciousness topics.

- There is also some online discussion of my stuff on [PSYCHE-D](#) (a mailing list devoted to consciousness). Here are a couple of threads on [The Conscious Mind](#) and on [Naturalism](#), and here are [my own postings](#) to PSYCHE-D.

Media discussion

Here are a few links to newspaper and magazine articles that discuss consciousness and mention my work. This sort of thing should be taken with a fairly large grain of salt.

- ["Can Machines Think?"](#)
-*Time* (Robert Wright), April 5, 1996
- ["Consciousness is Still Baffling to Experts of All Stripes"](#)
-*New York Times* (Sandra Blakeslee), April 16, 1996
- ["How to Make a Soul"](#)
-*Wired* (UK edition) (Andrew Brown), August 1996
- ["Zombies, Dolphins, and Blindsight"](#)
-*New Scientist*, June 1996.
- ["Can Science Explain Consciousness?"](#)
-*Scientific American* (John Horgan), July 1995
- ["Unraveling the Riddle of Identity"](#)
-*Los Angeles Times*, September 16, 1996
- ["Expert Says Consciousness is Basic Question of Reality"](#)
-*Yale Daily News*, September 18, 1996
- ["Untangling the Knot of Consciousness"](#)
-*UCSC Review* (Barbara McKenna), Winter 1997
- ["Trying to Explain our Consciousness"](#)
-*The Age*, March 4, 1997
- ["Consciousness Studies: From Stream to Flood"](#)
-*New York Times* (James Gorman), April 29, 1997.

Go to:

- [David Chalmers' home page](#)
- [Responses to articles on my work](#)
- [Reviews of *The Conscious Mind*](#)

Consciousness and its Place in Nature

David J. Chalmers

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

1 Introduction[*]

*[[This paper is an overview of issues concerning the metaphysics of consciousness. Much of the discussion in this paper (especially the first part) recapitulates discussion in Chalmers (1995; 1996; 1997), although it often takes a different form, and sometimes goes beyond the discussion there. I give a more detailed treatment of many of the issues discussed here in the works cited in the bibliography.]]

Consciousness fits uneasily into our conception of the natural world. On the most common conception of nature, the natural world is the physical world. But on the most common conception of consciousness, it is not easy to see how it could be part of the physical world. So it seems that to find a place for consciousness within the natural order, we must either revise our conception of consciousness, or revise our conception of nature.

In twentieth-century philosophy, this dilemma is posed most acutely in C. D. Broad's *The Mind and its Place in Nature* (Broad 1925). The phenomena of mind, for Broad, are the phenomena of consciousness. The central problem is that of locating mind with respect to the physical world. Broad's exhaustive discussion of the problem culminates in a taxonomy of seventeen different views of the mental-physical relation.[*] On Broad's taxonomy, a view might see the mental as nonexistent ("delusive"), as reducible, as emergent, or as a basic property of a substance (a "differentiating" attribute). The physical might be seen in one of the same four ways. So a four-by-four matrix of views results. (The seventeenth entry arises from Broad's division of the substance/substance view according to whether one substance or two is involved.) At the end, three views are left standing: those on which mentality is an emergent characteristic of either a physical substance or a neutral substance, where in the latter case, the physical might be either emergent or delusive.

*[[The taxonomy is in the final chapter, Chapter 14, of Broad's book (set out on pp. 607-11, and discussed until p. 650). The dramatization of Broad's taxonomy as a 4x4 matrix is illustrated on Andrew Chrucky's website devoted to Broad, at <http://www.ditext.com/broad/mpn14.html#t>.]]

In this paper I take my cue from Broad, approaching the problem of consciousness by a strategy of divide-and-conquer. I will not adopt Broad's categories: our understanding of the mind-body problem has

advanced in the last 75 years, and it would be nice to think that we have a better understanding of the crucial issues. On my view, the most important views on the metaphysics of consciousness can be divided almost exhaustively into six classes, which I will label "type A" through "type F." Three of these (A through C) involve broadly reductive views, seeing consciousness as a physical process that involves no expansion of a physical ontology. The other three (D through F) involve broadly nonreductive views, on which consciousness involves something irreducible in nature, and requires expansion or reconception of a physical ontology.

The discussion will be cast at an abstract level, giving an overview of the metaphysical landscape. Rather than engaging the empirical science of consciousness, or detailed philosophical theories of consciousness, I will be examining some general classes into which theories of consciousness might fall. I will not pretend to be neutral in this discussion. I think that each of the reductive views is incorrect, while each of the nonreductive views holds some promise. So the first part of this paper can be seen as an extended argument against reductive views of consciousness, while the second part can be seen as an investigation of where we go from there.

2 The Problem

The word 'consciousness' is used in many different ways. It is sometimes used for the ability to discriminate stimuli, or to report information, or to monitor internal states, or to control behavior. We can think of these phenomena as posing the "easy problems" of consciousness. These are important phenomena, and there is much that is not understood about them, but the problems of explaining them have the character of puzzles rather than mysteries. There seems to be no deep problem in principle with the idea that a physical system could be "conscious" in these senses, and there is no obvious obstacle to an eventual explanation of these phenomena in neurobiological or computational terms.

The hard problem of consciousness is the problem of experience. Humans beings have subjective experience: there is something it is like to be them. We can say that a being is conscious in this sense — or is phenomenally conscious, as it is sometimes put — when there is something it is like to be that being. A mental state is conscious when there is something it is like to be in that state. Conscious states include states of perceptual experience, bodily sensation, mental imagery, emotional experience, occurrent thought, and more. There is something it is like to see a vivid green, to feel a sharp pain, to visualize the Eiffel tower, to feel a deep regret, and to think that one is late. Each of these states has a *phenomenal character*, with *phenomenal properties* (or *qualia*) characterizing what it is like to be in the state.[*]

*[[On my usage, qualia are simply those properties that characterize conscious states according to what it is like to have them. The definition does not build in any further substantive requirements, such as the requirement that qualia are intrinsic or nonintentional. If qualia are intrinsic or nonintentional, this will be a substantive rather than a definitional point (so the claim that the properties of consciousness are non-intrinsic or that they are wholly intentional should not be taken to entail that there are no qualia). Phenomenal properties can also be taken to be properties of individuals (e.g., people) rather than of mental states, characterizing aspects of what it is like to be them at a given time; the difference will not matter much for present purposes.]]

There is no question that experience is closely associated with physical processes in systems such as brains. It seems that physical processes give rise to experience, at least in the sense that producing a physical system (such as a brain) with the right physical properties inevitably yields corresponding states of experience. But how and why do physical processes give rise to experience? Why do not these processes take place "in the dark," without any accompanying states of experience? This is the central mystery of consciousness.

What makes the easy problems easy? For these problems, the task is to explain certain behavioral or cognitive functions: that is, to explain how some causal role is played in the cognitive system, ultimately in the production of behavior. To explain the performance of such a function, one need only specify a mechanism that plays the relevant role. And there is good reason to believe that neural or computational mechanisms can play those roles.

What makes the hard problem hard? Here, the task is not to explain behavioral and cognitive functions: even once one has an explanation of all the relevant functions in the vicinity of consciousness — discrimination, integration, access, report, control — there may still remain a further question: why is the performance of these functions accompanied by experience? Because of this, the hard problem seems to be a different sort of problem, requiring a different sort of solution.

A solution to the hard problem would involve an account of the relation between physical processes and consciousness, explaining on the basis of natural principles how and why it is that physical processes are associated with states of experience. A *reductive explanation* of consciousness will explain this wholly on the basis of physical principles that do not themselves make any appeal to consciousness.[*] A *materialist* (or physicalist) solution will be a solution on which consciousness is itself seen as a physical process. A *nonmaterialist* (or nonphysicalist) solution will be a solution on which consciousness is seen as nonphysical (even if closely associated with physical processes). A *nonreductive* solution will be one on which consciousness (or principles involving consciousness) is admitted as a basic part of the explanation.

*[[Note that I use 'reductive' in a broader sense than it is sometimes used. Reductive explanation requires only that a high-level phenomena can be explained wholly in terms of low-level phenomena. This is compatible with the "multiple realizability" of high-level phenomena in low-level phenomena. For example, there may be many different ways in which digestion could be realized in a physiological system, but one can nevertheless reductively explain a system's digestion in terms of underlying physiology. Another subtlety concerns the possibility of a view on which consciousness can be explained in terms of principles which do not make appeal to consciousness but cannot themselves be physically explained. The definitions above count such a view as neither reductive nor nonreductive. It could reasonably be classified either way, but I will generally assimilate it with the nonreductive class.]]

It is natural to hope that there will be a materialist solution to the hard problem and a reductive explanation of consciousness, just as there have been reductive explanations of many other phenomena in many other domains. But consciousness seems to resist materialist explanation in a way that other phenomena do not. This resistance can be encapsulated in three related arguments against materialism, summarized in what follows.

3 Arguments against Materialism

3.1 The Explanatory Argument[*]

*[[A version of the explanatory argument as formulated here is given in Chalmers 1995. For related considerations about explanation, see Levine 1983 on the "explanatory gap" and Nagel 1974. See also the papers in Shear 1997.]]

The first argument is grounded in the difference between the easy problems and the hard problem, as characterized above: the easy problems concern the explanation of behavioral and cognitive functions, but the hard problem does not. One can argue that by the character of physical explanation, physical accounts explain *only* structure and function, where the relevant structures are spatiotemporal structures, and the relevant functions are causal roles in the production of a system's behavior. And one can argue as above that explaining structures and functions does not suffice to explain consciousness. If so, no physical account can explain consciousness.

We can call this the *explanatory argument*:

- (1) Physical accounts explain at most structure and function.
- (2) Explaining structure and function does not suffice to explain consciousness; so
-
- (3) No physical account can explain consciousness.

If this is right, then while physical accounts can solve the easy problems (which involve only explaining functions), something more is needed to solve the hard problem. It would seem that no reductive explanation of consciousness could succeed. And if we add the premise that what cannot be physically explained is not itself physical (this can be considered an additional final step of the explanatory argument), then materialism about consciousness is false, and the natural world contains more than the physical world.

Of course this sort of argument is controversial. But before examining various ways of responding, it is useful to examine two closely related arguments that also aim to establish that materialism about consciousness is false.

3.2 The Conceivability Argument.[*]

*[[Versions of the conceivability argument are put forward by Bealer 1994, Campbell 1970, Chalmers 1996, Kirk 1974, and Kripke 1980, among others. Important predecessors include Descartes' conceivability argument about disembodiment, and Leibniz's "mill" argument.]]

According to this argument, it is conceivable that there be a system that is physically identical to a conscious being, but that lacks at least some of that being's conscious states. Such a system might be a *zombie*: a system that is physically identical to a conscious being but that lacks consciousness entirely. It might also be an *invert*, with some of the original being's experiences replaced by different experiences, or a *partial zombie*, with some experiences absent, or a combination thereof. These systems will look identical to a normal conscious being from the third-person perspective: in particular, their brain processes will be molecule-for-molecule identical with the original, and their behavior will be indistinguishable. But things will be different from the first-person point of view. What it is like to be an invert or a partial zombie will differ from what it is like to be the original being. And there is nothing it is like to be a zombie.

There is little reason to believe that zombies exist in the actual world. But many hold that they are at least conceivable: we can coherently imagine zombies, and there is no contradiction in the idea that reveals itself even on reflection. As an extension of the idea, many hold that the same goes for a *zombie world*: a universe physically identical to ours, but in which there is no consciousness. Something similar applies to inverts and other duplicates.

From the conceivability of zombies, proponents of the argument infer their *metaphysical possibility*. Zombies are probably not naturally possible: they probably cannot exist in our world, with its laws of nature. But the argument holds that zombies *could have* existed, perhaps in a very different sort of universe. For example, it is sometimes suggested that God could have created a zombie world, if he had so chosen. From here, it is inferred that consciousness must be nonphysical. If there is a metaphysically possible universe that is physically identical to ours but that lacks consciousness, then consciousness must be a further, nonphysical component of our universe. If God could have created a zombie world, then (as Kripke puts it) after creating the physical processes in our world, he had to do more work to ensure that it contained consciousness.

We can put the argument, in its simplest form, as follows:

- (1) It is conceivable that there be zombies
- (2) If it is conceivable that there be zombies, it is metaphysically possible that there be zombies.
- (3) If it is metaphysically possible that there be zombies, then consciousness is nonphysical.

(4) Consciousness is nonphysical.

A somewhat more general and precise version of the argument appeals to P, the conjunction of all microphysical truths about the universe, and Q, an arbitrary phenomenal truth about the universe. (Here ' $\wedge$ ' represents 'and' and ' $\neg$ ' represents 'not'.)

- (1) It is conceivable that $P \wedge \neg Q$.
 - (2) If it is conceivable that $P \wedge \neg Q$, it is metaphysically possible that $P \wedge \neg Q$.
 - (3) If it is metaphysically possible that $P \wedge \neg Q$, then materialism is false.
-
- (4) Materialism is false.

3.3 The Knowledge Argument[*]

*[[Sources for the knowledge argument include Jackson 1982, Maxwell 1968, Nagel 1974, and others. Predecessors of the argument are present in Broad's discussion of a "mathematical archangel" who cannot deduce the smell of ammonia from physical facts (Broad 1925, pp. 70-71), and Feigl's discussion of a "Martian superscientist" who cannot know what colors look like and what musical tones sound like (Feigl 1958/1967, pp. 64, 68, 140).]]

According to the knowledge argument, there are facts about consciousness that are not deducible from physical facts. Someone could know all the physical facts, be a perfect reasoner, and still be unable to know all the facts about consciousness on that basis.

Frank Jackson's canonical version of the argument provides a vivid illustration. On this version, Mary is a neuroscientist who knows everything there is to know about the physical processes relevant to color vision. But Mary has been brought up in a black-and-white room (on an alternative version, she is colorblind[*]) and has never experienced red. Despite all her knowledge, it seems that there is something very important about color vision that Mary does not know: she does not know what it is like to see red. Even complete physical knowledge and unrestricted powers of deduction do not enable her to know this. Later, if she comes to experience red for the first time, she will learn a new fact of which she was previously ignorant: she will learn what it is like to see red.

*[[This version of the thought-experiment has a real life exemplar in Knut Nordby, a Norwegian sensory biologist who is a rod monochromat (lacking cones in his retina for color vision), and who works on the physiology of color vision. See Nordby 1990.]]

Jackson's version of the argument can be put as follows (here the premises concern Mary's knowledge when she has not yet experienced red):

- (1) Mary knows all the physical facts.
- (2) Mary does not know all the facts

(3) The physical facts do not exhaust all the facts.

One can put the knowledge argument more generally:

(1) There are truths about consciousness that are not deducible from physical truths.

(2) If there are truths about consciousness that are not deducible from physical truths, then materialism is false.

(3) Materialism is false.

3.4 The Shape of the Arguments

These three sorts of argument are closely related. They all start by establishing an *epistemic gap* between the physical and phenomenal domains. Each denies a certain sort of close epistemic relation between the domains: a relation involving what we can know, or conceive, or explain. In particular, each of them denies a certain sort of *epistemic entailment* from physical truths P to the phenomenal truths Q: deducibility of Q from P, or explainability of Q in terms of P, or conceiving of Q upon reflective conceiving of P.

Perhaps the most basic sort of epistemic entailment is a priori entailment, or *implication*. On this notion, P implies Q when the material conditional ' $P \supset Q$ ' is a priori; that is, when a subject can know that if P is the case then Q is the case, with justification independent of experience. All of the three arguments above can be seen as making a case against an a priori entailment of Q by P. If a subject who knows only P cannot deduce that Q (as the knowledge argument suggests), or if one can rationally conceive of P without Q (as the conceivability argument suggests), then it seems that P does not imply Q. The explanatory argument can be seen as turning on the claim that an implication from P to Q would require a functional analysis of consciousness, and that the concept of consciousness is not a functional concept.

After establishing an epistemic gap, these arguments proceed by inferring an ontological gap, where ontology concerns the nature of things in the world. The conceivability argument infers from conceivability to metaphysical possibility; the knowledge argument infers from failure of deducibility to difference in facts; and the explanatory argument infers from failure of physical explanation to nonphysicality. One might say that these arguments infer from a failure of epistemic entailment to a failure of ontological entailment. The paradigmatic sort of ontological entailment is *necessitation*: P necessitates Q when the material conditional ' $P \supset Q$ ' is metaphysically necessary, or when it is metaphysically impossible for P to hold without Q holding. It is widely agreed that materialism requires

that P necessitates all truths (perhaps with minor qualifications). So if there are phenomenal truths Q that P does not necessitate, then materialism is false.

We might call of these arguments *epistemic arguments* against materialism. Epistemic arguments arguably descend from Descartes' arguments against materialism (although these have a slightly different form), and are given their first thorough airing in Broad's book, which contains elements of all three arguments above.[*] The general form of an epistemic argument against materialism is as follows:

*[[For limited versions of the conceivability argument and the explanatory argument, see Broad, pp. 614-15. For the knowledge argument, see pp. 70-72, where Broad argues that even a "mathematical archangel" could not deduce the smell of ammonia from microscopic knowledge of atoms. Broad is arguing against "mechanism", which is roughly equivalently to contemporary materialism. Perhaps the biggest lacuna in Broad's argument, to contemporary eyes, is any consideration of the possibility that there is an epistemic but not an ontological gap.]]

(1) There is an epistemic gap between physical and phenomenal truths.

(2) If there is an epistemic gap between physical and phenomenal truths, then there is an ontological gap, and materialism is false.

—

(3) Materialism is false.

Of course this way of looking at things oversimplifies matters, and abstracts away from the differences between the arguments.[*] The same goes for the precise analysis in terms of implication and necessitation. Nevertheless, this analysis provides a useful lens through which to see what the arguments in common, and through which to analyze various responses to the arguments.

*[[For a discussion of the relationship between the conceivability argument and the knowledge argument, see Chalmers 1996 and Chalmers 2002b.]]

There are roughly three ways that a materialist might resist the epistemic arguments. A type-A materialist denies that there is the relevant sort of epistemic gap. A type-B materialist accepts that there is an unclosable epistemic gap, but denies that there is an ontological gap. And a type-C materialist accepts that there is a deep epistemic gap, but holds that it will eventually be closed. In what follows, I discuss all three of these strategies.

4 Type-A Materialism

According to type-A materialism, there is no epistemic gap between physical and phenomenal truths; or at least, any apparent epistemic gap is easily closed. According to this view, it is not conceivable (at least on reflection) that there be duplicates of conscious beings that have absent or inverted conscious states. On

this view, there are no phenomenal truths of which Mary is ignorant in principle from inside her black-and-white room (when she leaves the room, she gains at most an ability). And on this view, on reflection there is no "hard problem" of explaining consciousness that remains once one has solved the easy problems of explaining the various cognitive, behavioral, and environmental functions.[*]

*[[Type-A materialists include Dennett 1991, Dretske 1995, Harman 1990, Lewis 1988, Rey 1995, and Ryle 1949.]]

Type-A materialism sometimes takes the form of eliminativism, holding that consciousness does not exist, and that there are no phenomenal truths. It sometimes takes the form of analytic functionalism or logical behaviorism, holding that consciousness exists, where the concept of "consciousness" is defined in wholly functional or behavioral terms (e.g., where to be conscious might be to have certain sorts of access to information, and/or certain sorts of dispositions to make verbal reports). For our purposes, the difference between these two views can be seen as terminological. Both agree that we are conscious in the sense of having the functional capacities of access, report, control, and the like; and they agree that we are not conscious in any further (nonfunctionally defined) sense. The analytic functionalist thinks that ordinary terms such as 'conscious' should be used in the first sort of sense (expressing a functional concept), while the eliminativist thinks that it should be used in the second. Beyond this terminological disagreement about the use of existing terms and concepts, the substance of the views is the same.

Some philosophers and scientists who do not explicitly embrace eliminativism, analytic functionalism, and the like are nevertheless recognizably type-A materialists. The characteristic feature of the type-A materialist is the view that on reflection there is nothing in the vicinity of consciousness that needs explaining over and above explaining the various functions: to explain these things is to explain everything in the vicinity that needs to be explained. The relevant functions may be quite subtle and complex, involving fine-grained capacities for access, self-monitoring, report, control, and their interaction, for example. They may also be taken to include all sorts of environmental relations. And the explanation of these functions will probably involve much neurobiological detail. So views that are put forward as rejecting functionalism on the grounds that it neglects biology or neglects the role of the environment may still be type-A views.

One might think that there is room in logical space for a view that denies even this sort of broadly functionalist view of consciousness, but still holds that there is no epistemic gap between physical and phenomenal truths. In practice, there appears to be little room for such a view, for reasons that I will discuss under type C, and there are few examples of such views in practice.[*] So I will take it for granted that a type-A view is one that holds that explaining the functions explains everything, and will class other views that hold that there is no unclosable epistemic gap under type C.

*[[Two specific views may be worth mentioning. (i) Some views (e.g., Dretske 1995) deny an epistemic gap while at the same time denying functionalism, by holding that consciousness involves not just functional role but also causal and historical relations to objects in the environment. I count these as type-A views: we can view the relevant relations as part of functional role, broadly construed, and exactly the same considerations arise. (ii) Some views (e.g., Stoljar 2001 and Strawson 2000) deny an epistemic gap not by functionally analyzing consciousness but by expanding our view of the physical base to include underlying intrinsic properties. These views are discussed under type F.]]

The obvious problem with type-A materialism is that it appears to deny the manifest. It is an uncontested truth that we have the various functional capacities of access, control, report, and the like, and these phenomena pose uncontested explananda (phenomena in need of explanation) for a science of consciousness. But in addition, it seems to be a further truth that we are conscious, and this phenomenon seems to pose a further explanandum. It is this explanandum that raises the interesting problems of consciousness. To flatly deny the further truth, or to deny without argument that there is a hard problem of consciousness over and above the easy problems, would be to make a highly counterintuitive claim that begs the important questions. This is not to say that highly counterintuitive claims are always false, but they need to be supported by extremely strong arguments. So the crucial question is: are there any compelling *arguments* for the claim that on reflection, explaining the functions explains everything?

Type-A materialists often argue by analogy. They point out that in other areas of science, we accept that explaining the various functions explains the phenomena, so we should accept the same here. In response, an opponent may well accept that in other domains, the functions are all we need to explain. In explaining life, for example, the only phenomena that present themselves as needing explanation are phenomena of adaptation, growth, metabolism, reproduction, and so on, and there is nothing else that even calls out for explanation. But the opponent holds that the case of consciousness is different and possibly unique, precisely because there is something else, phenomenal experience, that calls out for explanation. The type-A materialist must either deny even the appearance of a further explanandum, which seems to deny the obvious, or accept the apparent disanalogy and give further substantial arguments for why, contrary to appearances, only the functions need to be explained.

At this point, type-A materialists often press a different sort of analogy, holding that at various points in the past, thinkers held that there was an analogous epistemic gap for other phenomena, but that these turned out to be physically explained. For example, Dennett (1996) suggests that a vitalist might have held that there was a further "hard problem" of life over and above explaining the biological function, but that this would have been misguided.

On examining the cases, however, the analogies do not support the type-A materialist. Vitalists typically *accepted*, implicitly or explicitly, that the biological functions in question were what needed explaining. Their vitalism arose because they thought that the functions (adaptation, growth, reproduction, and so on) would not be physically explained. So this is quite different from the case of consciousness. The disanalogy is very clear in the case of Broad. Broad was a vitalist about life, holding that the functions would require a non-mechanical explanation. But at the same time, he held that in the case of life, unlike the case of consciousness, the only evidence we have for the phenomenon is behavioral, and that "being alive" means exhibiting certain sorts of behavior. Other vitalists were less explicit, but very few of them held that something more than the functions needed explaining (except consciousness itself, in some cases). If a vitalist had held this, the obvious reply would have been that there is no reason to believe in such an explanandum. So there is no analogy here.[*]

*[[In another analogy, Churchland (1996) suggests that someone in Goethe's time might have mounted analogous epistemic arguments against the reductive explanation of "luminescence." But on a close look, it is not hard to see that the only further explanandum that could have caused doubts here is the *experience* of seeing light (see Chalmers 1997). This

point is no help to the type-A materialist, since this explanandum remains unexplained.]]

So these arguments by analogy have no force for the type-A materialist. In other cases, it was always clear that structure and function exhausted the apparent explananda, apart from those tied directly to consciousness itself. So the type-A materialist needs to address the apparent further explanandum in the case of consciousness head on: either flatly denying it, or giving substantial arguments to dissolve it.

Some arguments for type-A materialists proceed indirectly, by pointing out the unsavory metaphysical or epistemological consequences of rejecting the view: e.g., that the rejection leads to dualism, or to problems involving knowledge of consciousness.[*] An opponent will either embrace the consequences or deny that they are consequences. As long as the consequences are not completely untenable, then for the type-A materialist to make progress, this sort of argument needs to be supplemented by a substantial direct argument against the further explanandum.

*[[For an argument from unsavory metaphysical consequences, see White 1986. For an argument from unsavory epistemological consequences, see Shoemaker 1975. The metaphysical consequences are addressed in the second half of this paper. The epistemological consequences are addressed in Chalmers 2002a.]]

Such direct arguments are surprisingly hard to find. Many arguments for type-A materialism end up presupposing the conclusion at crucial points. For example, it is sometimes argued (e.g., Rey 1995) that there is no reason to postulate qualia, since they are not needed to explain behavior; but this argument presupposes that only behavior needs explaining. The opponent will hold that qualia are an explanandum in their own right. Similarly, Dennett's use of "heterophenomenology" (verbal reports) as the primary data to ground his theory of consciousness (Dennett 1991) appears to rest on the assumption that these reports are what need explaining, or that the only "seemings" that need explaining are dispositions to react and report.

One way to argue for type-A materialism is to argue that there is some intermediate X such that (i) explaining functions suffices to explain X, and (ii) explaining X suffices to explain consciousness. One possible X here is *representation*: it is often held both that conscious states are representational states, representing things in the world, and that we can explain representation in functional terms. If so, it may seem to follow that we can explain consciousness in functional terms. On examination, though, this argument appeals to an ambiguity in the notion of representation. There is a notion of *functional representation*, on which P is represented roughly when a system responds to P and/or produces behavior appropriate for P. In this sense, explaining functioning may explain representation, but explaining representation does not explain consciousness. There is also a notion of *phenomenal representation*, on which P is represented roughly when a system has a conscious experience as if P. In this sense, explaining representation may explain consciousness, but explaining functioning does not explain representation. Either way, the epistemic gap between the functional and the phenomenal remains as wide as ever. Similar sorts of equivocation can be found with other X's that might be appealed to here, such as "perception" or "information."

Perhaps the most interesting arguments for type-A materialism are those that argue that we can give a

physical explanation of our *beliefs* about consciousness, such as the belief that we are conscious, the belief that consciousness is a further explanandum, and the belief that consciousness is nonphysical. From here it is argued that once we have explained the belief, we have done enough to explain, or to explain away, the phenomenon (e.g., Clark 2000, Dennett forthcoming). Here it is worth noting that this only works if the beliefs themselves are functionally analyzable; Chalmers (2002a) gives reason to deny this. But even if one accepts that beliefs are ultimately functional, this claim then reduces to the claim that explaining our dispositions to talk about consciousness (and the like) explains everything. An opponent will deny this claim: explaining the dispositions to report may remove the third-person warrant (based on observation of others) for accepting a further explanandum, but it does not remove the crucial first-person warrant (from one's own case). Still, this is a strategy that deserves extended discussion.

At a certain point, the debate between type-A materialists and their opponents usually comes down to intuition: most centrally, the intuition that consciousness (in a nonfunctionally defined sense) exists, or that there is something that needs to be explained (over and above explaining the functions). This claim does not gain its support from argument, but from a sort of observation, along with rebuttal of counterarguments. The intuition appears to be shared by the large majority of philosophers, scientists, and others; and it is so strong that to deny it, a type-A materialist needs exceptionally powerful arguments. The result is that even among materialists, type-A materialists are a distinct minority.

5 Type-B Materialism[*]

*[[Type-B materialists include Block and Stalnaker 1999, Hill 1997, Levine 1983, Loar 1990/1997, Lycan 1996, Papineau 1993, Perry 2001, and Tye 1995.]]

According to type-B materialism, there is an epistemic gap between the physical and phenomenal domains, but there is no ontological gap. According to this view, zombies and the like are conceivable, but they are not metaphysically possible. On this view, Mary is ignorant of some phenomenal truths from inside her room, but nevertheless these truths concern an underlying physical reality (when she leaves the room, she learns old facts in a new way). And on this view, while there is a hard problem distinct from the easy problems, it does not correspond to a distinct ontological domain.

The most common form of type-B materialism holds that phenomenal states can be *identified* with certain physical or functional states. This identity is held to be analogous in certain respects (although perhaps not in all respects) with the identity between water and H₂O, or between genes and DNA.[*] These identities are not derived through conceptual analysis, but are discovered empirically: the concept *water* is different from the concept *H₂O*, but they are found to refer to the same thing in nature. On the type-B view, something similar applies to consciousness: the concept of consciousness is distinct from any physical or functional concepts, but we may discover empirically that these refer to the same thing in nature. In this way, we can explain why there is an epistemic gap between the physical and phenomenal domains, while denying any ontological gap. This yields the attractive possibility that we can acknowledge the deep epistemic problems of consciousness while retaining a materialist worldview.

*[[In certain respects, where type-A materialism can be seen as deriving from the logical behaviorism of Ryle and Carnap, type-B materialism can be seen as deriving from the identity theory of Place and Smart. The matter is complicated, however, by the fact that the early identity-theorists advocated "topic-neutral" (functional) analyses of phenomenal properties, suggesting an underlying type-A materialism.]]

Although such a view is attractive, it faces immediate difficulties. These difficulties stem from the fact that the character of the epistemic gap with consciousness seems to differ from that of epistemic gaps in other domains. For a start, there do not seem to be analogs of the epistemic arguments above in the cases of water, genes, and so on. To explain genes, we merely have to explain why systems function a certain way in transmitting hereditary characteristics; to explain water, we have to explain why a substance has a certain objective structure and behavior. Given a complete physical description of the world, Mary would be able to deduce all the relevant truths about water and about genes, by deducing which systems have the appropriate structure and function. Finally, it seems that we cannot coherently conceive of a world physically identical to our own, in which there is no water, or in which there are no genes. So there is no epistemic gap between the *complete* physical truth about the world and the truth about water and genes that is analogous to the epistemic gap with consciousness.

(Except, perhaps, for epistemic gaps that derive from the epistemic gap for consciousness. For example, perhaps Mary could not deduce or explain the perceptual *appearance* of water from the physical truth about the world. But this would just be another instance of the problem we are concerned with, and so cannot help the type-B materialist.)

So it seems that there is something unique about the case of consciousness. We can put this by saying that while the identity between genes and DNA is empirical, it is not *epistemically primitive*: the identity is itself deducible from the complete physical truth about the world. By contrast, the type-B materialist must hold that the identification between consciousness and physical or functional states is epistemically primitive: the identity is not deducible from the complete physical truth. (If it were deducible, type-A materialism would be true instead.) So the identity between consciousness and a physical state will be a sort of primitive principle in one's theory of the world.

Here, one might suggest that something has gone wrong. Elsewhere, the only sort of place that one finds this sort of primitive principle is in the fundamental laws of physics. Indeed, it is often held that this sort of primitiveness — the inability to be deduced from more basic principles — is the mark of a fundamental law of nature. In effect, the type-B materialist recognizes a principle that has the epistemic status of a fundamental law, but gives it the ontological status of an identity. An opponent will hold that this move is more akin to theft than to honest toil: elsewhere, identifications are grounded in explanations, and primitive principles are acknowledged as fundamental laws.

It is natural to suggest that the same should apply here. If one acknowledges the epistemically primitive connection between physical states and consciousness as a fundamental law, it will follow that consciousness is distinct from any physical property, since fundamental laws always connect distinct properties. So the usual standard will lead to one of the nonreductive views discussed in the second half of this paper. By contrast, the type-B materialist takes an observed connection between physical and

phenomenal states, unexplainable in more basic terms, and suggests that it is an identity. This suggestion is made largely in order to preserve a prior commitment to materialism. Unless there is an independent case for primitive identities, the suggestion will seem at best ad hoc and mysterious, and at worst incoherent.

A type-B materialist might respond in various ways. First, some (e.g., Papineau 1993) suggest that identities do not *need* to be explained, so are always primitive. But we have seen that identities in other domains can at least be *deduced* from more basic truths, and so are not primitive in the relevant sense. Second, some (e.g., Block and Stalnaker 1999) suggest that even truths involving water and genes cannot be deduced from underlying physical truths. This matter is too complex to go into here (see Chalmers and Jackson 2001 for a response[*]), but one can note that the epistemic arguments outlined at the beginning suggest a very strong disanalogy between consciousness and other cases. Third, some (e.g., Loar 1990/1997) acknowledge that identities involving consciousness are unlike other identities by being epistemically primitive, but seek to explain this uniqueness by appealing to unique features of the concept of consciousness. This response is perhaps the most interesting, and I will return to it.

*[[Block and Stalnaker (1999) argue against deducibility in part by arguing that there is usually no explicit conceptual analysis of high-level terms such as 'water' in microphysical terms, or in any other terms that could ground an a priori entailment from microphysical truths to truths about water. In response, Chalmers and Jackson (2001) argue that explicit conceptual analyses are not required for a priori entailments, and that there is good reason to believe that such entailments exist in these cases.]]

There is another line that a type-B materialist can take. One can first note that an *identity* between consciousness and physical states is not strictly required for a materialist position. Rather, one can plausibly hold that materialism about consciousness simply requires that physical states *necessitate* phenomenal states, in that it is metaphysically impossible for the physical states to be present while the phenomenal states are absent or different. That is, materialism requires that entailments $P \supset Q$ be necessary, where P is the complete physical truth about the world and Q is an arbitrary phenomenal truth.

At this point, a type-B materialist can naturally appeal to the work of Kripke (1980), which suggests that some truths are necessarily true without being a priori. For example, Kripke suggests that 'water is H₂O' is necessary — true in all possible worlds — but not knowable a priori. Here, a type-B materialist can suggest that $P \supset Q$ may be a Kripkean a posteriori necessity, like 'water is H₂O' (though it should be noted that Kripke himself denies this claim). If so, then we would *expect* there to be an epistemic gap, since there is no a priori entailment from P to Q, but at the same time there will be no ontological gap. In this way, Kripke's work can seem to be just what the type-B materialist needs.

Here, some of the issues that arose previously arise again. One can argue that in other domains, necessities are not epistemically primitive. The necessary connection between water and H₂O may be a posteriori, but it can itself be deduced from a complete physical description of the world (one can deduce that water is identical to H₂O, from which it follows that water is necessarily H₂O). The same applies to the other necessities that Kripke discusses. By contrast, the type-B materialist must hold that the

connection between physical states and consciousness is epistemically primitive, in that it cannot be deduced from the complete physical truth about the world. Again, one can suggest that this sort of primitive necessary connection is mysterious and ad hoc, and that the connection should instead be viewed as a fundamental law of nature.

I will discuss further problems with these necessities in the next section. But here, it is worth noting that there is a sense in which any type-B materialist position gives up on reductive explanation. Even if type-B materialism is true, we cannot give consciousness the same sort of explanation that we give genes and like, in purely physical terms. Rather, our explanation will always require explanatorily primitive principles to bridge the gap from the physical to the phenomenal. The *explanatory* structure of a theory of consciousness, on such a view, will be very much unlike that of a materialist theory in other domains, and very much like the explanatory structure of the nonreductive theories described below. By labeling these principles identities or necessities rather than laws, the view may preserve the letter of materialism; but by requiring primitive bridging principles, it sacrifices much of materialism's spirit.

6 The Two-Dimensional Argument against Type-B Materialism

As discussed above, the type-B materialist holds that zombie worlds and the like are conceivable (there is no contradiction in $P \wedge \neg Q$) but are not metaphysically possible. That is, $P \supset Q$ is held to be an a posteriori necessity, akin to such a posteriori necessities as 'water is H_2O '. We can analyze this position in more depth by taking a closer look at the Kripkean cases of a posteriori necessity. This material is somewhat technical (hence the separate section) and can be skipped if necessary on a first reading.

It is often said that in Kripkean cases, conceivability does not entail possibility: it is conceivable that water is not H_2O (in that it is coherent to suppose that water is not H_2O), but it is not possible that water is not H_2O . But at the same time, it seems that there is *some* possibility in the vicinity of what one conceives. When one conceives that water is not H_2O , one conceives of a world W (the XYZ-world) in which the watery liquid in the oceans is not H_2O , but XYZ, say. There is no reason to doubt that the XYZ-world is metaphysically possible. If Kripke is correct, the XYZ-world is not correctly described as one in which water is H_2O . Nevertheless, this world is relevant to the truth of 'water is H_2O ' in a slightly different way, which can be brought out as follows.

One can say that the XYZ-world could *turn out* to be actual, in that for all we know a priori, the actual world is just like the XYZ-world. And one can say that *if* the XYZ-world turns out to be actual, it will turn out that water is XYZ. Similarly: if we hypothesize that the XYZ-world is actual, we should rationally conclude on that basis that water is not H_2O . That is, there is a deep *epistemic* connection between the XYZ-world and 'water is not H_2O '. Even Kripke allows that it is *epistemically possible* that water is not H_2O (in the broad sense that this is not ruled out a priori). It seems that the epistemic possibility that the XYZ-world is actual is a specific instance of the epistemic possibility that water is not H_2O .

Here, we adopt a special attitude to a world W . We think of W as an epistemic possibility: as a way the world might actually be. When we do this, we consider W *as actual*. When we think of W as actual, it may make a given sentence S true or false. For example, when thinking of the XYZ-world as actual, it makes 'water is not H_2O ' true. This is brought out in the intuitive judgment that if W turns out to be actual, it will turn out that water is not H_2O , and that the epistemic possibility that W is actual is an instance of the epistemic possibility that water is H_2O .

By contrast, one can also consider a world W *as counterfactual*. When we do this, we acknowledge that the character of the actual world is already fixed, and we think of W as a counterfactual way things might have been but are not. If Kripke is right, then if the watery stuff *had been* XYZ, XYZ would nevertheless not have been water. So when we consider the XYZ-world as counterfactual, it does not make 'water is not H_2O ' true. Considered as counterfactual, we describe the XYZ-world in light of the actual-world fact that water is H_2O , and we conclude that XYZ is not water but merely watery stuff. These results do not conflict: they simply involve two different ways of considering and describing possible worlds. Kripke's claims consider *counterfactual* evaluation of worlds, whereas the claims in the previous paragraph concern the *epistemic* evaluation of worlds.

One can formalize this using *two-dimensional semantics* (see Chalmers (this volume, chapter 56)).[*] We can say that if W considered as actual makes S true, then W *verifies* S , and that if W considered as counterfactual makes S true, then W *satisfies* S . Verification involves the epistemic evaluation of worlds, whereas satisfaction involves the counterfactual evaluation of worlds. Correspondingly, we can associate S with different *intensions*, or functions from worlds to truth values. The *primary* (or epistemic) intension of S is a function that is true at a world W iff W verifies S , and the *secondary* (or subjunctive) intension is a function that is true at a world W if W satisfies S . For example, where S is 'water is not H_2O ', and W is the XYZ-world, we can say that W verifies S but W does not satisfy S ; and we can say that the primary intension of S is true at W , but the secondary intension of S is false at W .

*[[Two-dimensional semantic frameworks originate in work of Kaplan (1989), Stalnaker (1978), and Evans (1979). The version used in these arguments is somewhat different: for discussion of the differences, see Chalmers (forthcoming).]]

With this in mind, one can suggest that when a statement S is conceivable — that is, when its truth cannot be ruled out a priori — then there is some world that verifies S , or equivalently, there is some world at which S 's primary intension is true. This makes intuitive sense: when S is conceivable, S represents an epistemic possibility. It is natural to suggest that corresponding to these epistemic possibilities are specific worlds W , such that when these are considered *as epistemic possibilities*, they verify S . That is, W is such that intuitively, if W turns out to be actual, it would turn out that S .

This model seems to fit all of Kripke's cases. For example, Kripke holds that it is an a posteriori necessity that heat is the motion of molecules. So it is conceivable in the relevant sense that heat is not the motion of molecules. Corresponding to this conceivable scenario is a world W in which heat sensations are caused by something other than the motion of molecules. W represents an epistemic possibility: and we can say that if W turns out to be actual, it will turn out that heat is not the motion of molecules. The same

goes in many other cases. The moral is that these Kripkean phenomena involve two different ways of thinking of possible worlds, with just one underlying space of worlds.

If this principle is applied to the case of type-B materialism, trouble immediately arises. As before, let P be the complete physical truth about the world, and let Q be a phenomenal truth. Let us say that S is conceivable when the truth of S is not ruled out a priori. Then one can mount an argument as follows:[*]

*[[This is a slightly more formal version of an argument in Chalmers 1996 (pp. 131-36). It is quite closely related to Kripke's modal argument against the identity theory, though different in some important respects. The central premise 2 can be seen as a way of formalizing Kripke's claim that where there is "apparent contingency", there is some misdescribed possibility in the background. The argument can also be seen as a way of formalizing a version of the "dual property" objection attributed to Max Black by Smart 1959, and developed by Jackson 1979 and White 1986. Related applications of the two-dimensional framework to questions about materialism are given by Jackson 1994 and Lewis 1994.]]

(1) $P \wedge \neg Q$ is conceivable

(2) If $P \wedge \neg Q$ is conceivable, a world verifies $P \wedge \neg Q$.

(3) If a world verifies $P \wedge \neg Q$, then a world satisfies $P \wedge \neg Q$ or type-F monism is true.

(4) If a world satisfies $P \wedge \neg Q$, materialism is false.

—

(5) Materialism is false or type-F monism is true.

The type-B materialist grants premise (1): to deny this would be to accept type-A materialism. Premise (2) is an instance of the general principle discussed above. Premise (4) can be taken as definitive of materialism. As for premise (3): in general one cannot immediately move from a world verifying S to a world satisfying S , as the case of 'water is H_2O ' (and the XYZ-world) suggests. But in the case of $P \wedge \neg Q$, a little reflection on the nature of P and Q takes us in that direction, as follows.

First, Q . Here, it is plausible that if W verifies 'there is consciousness', then W satisfies 'there is consciousness', and vice versa. This corresponds to the Kripkean point that in the case of consciousness, there is no distinction analogous to that between water itself and mere watery stuff. To put it intuitively, if W verifies 'there is consciousness', it contains something that at least *feels* conscious, and if something *feels* conscious, it *is* conscious. One can hold more generally that the primary and secondary intensions of our core phenomenal concepts are the same (see Chalmers 2002a). It follows that if world W verifies $\neg Q$, W satisfies $\neg Q$. (This claim is not required for the argument to go through, but it is plausible and makes things more straightforward.)

Second, P . A type-B materialist might seek to evade the argument by arguing that while W verifies P , it

does not satisfy P. On reflection, the only way this might work is as follows. If a world verifies P, it must have at least the *structure* of the actual physical world. The only reason why W might not satisfy P is that it lacks the intrinsic properties underlying this structure in the actual world. (On this view, the primary intension of a physical concept picks out whatever property plays a certain role in a given world, and the secondary intension picks out the actual intrinsic property across all worlds.) If this difference in W is responsible for the absence of consciousness in W, it follows that consciousness in the actual world is not necessitated by the structural aspects of physics, but by its underlying intrinsic nature. This is precisely the position I call type-F monism, or "panprotopsyism." Type-F monism is an interesting and important position, but it is much more radical than type-B materialism as usually conceived, and I count it as a different position. I will defer discussion of the reasoning and of the resulting position until then.

It follows that premise (4) is correct. If a world verifies $P \wedge \neg Q$, then either a world satisfies $P \wedge \neg Q$, or type-F monism is true. Setting aside type-F monism for now, it follows that the physical truth about our world does not necessitate the phenomenal truth, and materialism is false.

This conclusion is in effect a consequence of (i) the claim that $P \wedge \neg Q$ is conceivable (in the relevant sense), (ii) the claim that when S is conceivable, there is a world that verifies S, and (iii) some straightforward reasoning. A materialist might respond by denying (i), but that is simply to deny the relevant epistemic gap between the physical and the phenomenal, and so to deny type-B materialism. I think there is little promise for the type-B materialist in denying the reasoning involved in (iii). So the only hope for the type-B materialist is to deny the central thesis (ii).[*]

*[[I have passed over a few subtleties here. One concerns the role of indexicals: to handle claims such as 'I am here', primary intensions are defined over *centered worlds*: worlds with a marked individual and time, corresponding to indexical "locating information" about one's position in the world. This change does not help the type-B materialist, however. Even if we supplement P with indexical locating information I (e.g., telling Mary about her location in the world), there is as much of an epistemic gap with Q as ever; so $P \wedge I \wedge \neg Q$ is conceivable. And given that there is a centered world that verifies $P \wedge I \wedge \neg Q$, one can see as above that either there is a world satisfying $P \wedge \neg Q$, or type-F monism is true.]]

To do this, a type-B materialist could deny the coherence of the distinction between verification and satisfaction, or accept that the distinction is coherent but deny that thesis (ii) holds even in the standard Kripkean cases, or accept that thesis (ii) holds in the standard Kripkean cases but deny that it holds in the special case of consciousness. The first two options deserve exploration, but I think they are ultimately unpromising, as the distinction and the thesis appear to fit the Kripkean phenomena very well. Ultimately, I think a type-B materialist must hold that the case of consciousness is special, and that the thesis that holds elsewhere fails here.

On this view, the a posteriori necessities connecting the physical and phenomenal domains are much stronger than those in other domains, in that they are verified by all worlds. Elsewhere, I have called these unusual a posteriori necessities *strong necessities*, and have argued that there is no good reason to believe they exist. As with explanatorily primitive identities, they appear to be primitive facts postulated in an ad hoc way, largely in order to save a theory, with no support from cases elsewhere. Further, one can argue that this view leads to an underlying *modal dualism*, with independent primitive domains of logical and

metaphysical possibility; and one can argue that this is unacceptable.

Perhaps the most interesting response from a type-B materialist is to acknowledge that strong necessities are unique to the case of consciousness, and to try to explain this uniqueness in terms of special features of our conceptual system. For example, Christopher Hill (1997) has argued that one can predict the epistemic gap in the case of consciousness from the fact that physical concepts and phenomenal concepts have different conceptual roles. Brian Loar (1990/1997) has appealed to the claim that phenomenal concepts are recognitional concepts that lack contingent modes of presentation. Joseph Levine (1998) has argued that phenomenal concepts have nonascriptive modes of presentation. In response, I have argued (Chalmers 1999) that these responses do not work, and that there are systematic reasons why they cannot work.[*] But it is likely that further attempts in this direction will be forthcoming. This remains one of the key areas of debate on the metaphysics of consciousness.

*[[Hill (1997) tries to explain away our modal intuitions about consciousness in cognitive terms. Chalmers (1999) responds that any modal intuition might be explained in cognitive terms (a similar argument could "explain away" our intuition that there might be red squares), but that this has no tendency to suggest that the intuition is incorrect. If such an account tells us that modal intuitions about consciousness are unreliable, the same goes for all modal intuitions. What is really needed is not an explanation of our modal intuitions about consciousness, but an explanation of why these intuitions in particular should be unreliable.]

Loar (1990/1997) attempts to provide such an explanation in terms of the unique features of phenomenal concepts. He suggests that (1) phenomenal concepts are recognitional concepts ("*that* sort of thing"); that (2) like other recognitional concepts, they can corefer with physical concepts that are cognitively distinct; and that (3) unlike other recognitional concepts, they lack contingent modes of presentation (i.e., their primary and secondary intensions coincide). If (2) and (3) both hold (and if we assume that physical concepts also lack contingent modes of presentation), then a phenomenal-physical identity will be a strong necessity in the sense above. In response, Chalmers (1999) argues that (2) and (3) cannot both hold. The coreference of other recognitional concepts with theoretical concepts is *grounded* in their contingent modes of presentation; in the absence of such modes of presentation, there is no reason to think that these concepts can corefer. So accepting (3) undercuts any support for (2). Chalmers (1999) also argues that by assuming that physical properties can have phenomenal modes of presentation noncontingently, Loar's account is in effect presupposing rather than explaining the relevant strong necessities.]]

Overall, my own view is that there is little reason to think that explanatorily primitive identities or strong necessities exist. There is no good *independent* reason to believe in them: the best reason to postulate them is to save materialism, but in the context of a debate over whether materialism is true this reasoning is unconvincing, especially if there are viable alternatives. Nevertheless, further investigation into the key issues underlying this debate is likely to be philosophically fruitful.

7 Type-C Materialism

According to type-C materialism, there is a deep epistemic gap between the physical and phenomenal domains, but it is closable in principle. On this view, zombies and the like are conceivable for us now, but they will not be conceivable in the limit. On this view, it currently seems that Mary lacks information about the phenomenal, but in the limit there would be no information that she lacks. And on this view, while we cannot see now how to solve the hard problem in physical terms, the problem is solvable in

principle.

This view is initially very attractive. It seems to acknowledge the deep explanatory gap with which we seem to be faced, while at the same time allowing that the apparent gap may be due to our own limitations. There are different versions of the view. Nagel (1974) has suggested that just as the pre-Socratics could not have understood how matter could be energy, we cannot understand how consciousness could be physical, but a conceptual revolution might allow the relevant understanding. Churchland (1997) suggests that even if we cannot now imagine how consciousness could be a physical process, that is simply a psychological limitation on our part that further progress in science will overcome. Van Gulick (1993) suggests that conceivability arguments are question-begging, since once we have a good explanation of consciousness, zombies and the like will no longer be conceivable. McGinn (1989) has suggested that the problem may be unsolvable by humans due to deep limitations in our cognitive abilities, but that it nevertheless has a solution in principle.

One way to put the view is as follows. Zombies and the like are *prima facie* conceivable (for us now, with our current cognitive processes), but they are not *ideally* conceivable (under idealized rational reflection). Or we could say: phenomenal truths are deducible in principle from physical truths, but the deducibility is akin to that of a complex truth of mathematics: it is accessible in principle (perhaps accessible a priori), but is not accessible to us now, perhaps because the reasoning required is currently beyond us, or perhaps because we do not currently grasp all the required physical truths. If this is so, then there will appear to us that there is a gap between physical processes and consciousness, but there will be no gap in nature.

Despite its appeal, I think that the type-C view is inherently unstable. Upon examination, it turns out either to be untenable, or to collapse into one of the other views on the table. In particular, it seems that the view must collapse into a version of type-A materialism, type-B materialism, type-D dualism, or type-F monism, and so is not ultimately a distinct option.

One way to hold that the epistemic gap might be closed in the limit is to hold that in the limit, we will see that explaining the functions explains everything, and that there is no further explanandum. It is at least coherent to hold that we currently suffer from some sort of conceptual confusion or unclarity that leads us to believe that there is a further explanandum, and that this situation could be cleared up by better reasoning. I will count this position as a version of type-A materialism, not type-C materialism: it is obviously closely related to standard type-A materialism (the main difference is whether we have yet had the relevant insight), and the same issues arise. Like standard type-A materialism, this view ultimately stands or falls with the strength of (actual and potential) first-order arguments that dissolve any apparent further explanandum.

Once type-A materialism is set aside, the potential options for closing the epistemic gap are highly constrained. These constraints are grounded in the nature of physical concepts, and in the nature of the concept of consciousness. The basic problem has already been mentioned. First: Physical descriptions of the world characterize the world in terms of structure and dynamics. Second: From truths about structure and dynamics, one can deduce only further truths about structure and dynamics. And third: truths about consciousness are not truths about structure and dynamics. But we can take these steps one at a time.

First: A microphysical description of the world specifies a distribution of particles, fields, and waves in space and time. These basic systems are characterized by their spatiotemporal properties, and properties such as mass, charge, and quantum wavefunction state. These latter properties are ultimately defined in terms of spaces of states that have a certain abstract structure (e.g., the space of continuously varying real quantities, or of Hilbert space states), such that the states play a certain causal role with respect to other states. We can subsume spatiotemporal descriptions and descriptions in terms of properties in these formal spaces under the rubric of *structural* descriptions. The state of these systems can change over time in accord with dynamic principles defined over the relevant properties. The result is a description of the world in terms of its underlying spatiotemporal and formal structure, and dynamic evolution over this structure.

Some type-C materialists hold we do not yet have a complete physics, so we cannot know what such a physics might explain. But here we do not need to have a complete physics: we simply need the claim that physical descriptions are in terms of structure and dynamics. This point is general across physical theories. Such novel theories as relativity, quantum mechanics, and the like may introduce new structures, and new dynamics over those structures, but the general point (and the gap with consciousness) remains.

A type-C materialist might hold that there could be new physical theories that go beyond structure and dynamics. But given the character of physical explanation, it is unclear what sort of theory this could be. Novel physical properties are postulated for their potential in explaining existing physical phenomena, themselves characterized in terms of structure and dynamics, and it seems that structure and dynamics always suffices here. One possibility is that instead of postulating novel properties, physics might end up appealing to consciousness itself, in the way that some theorists hold that quantum mechanics does. This possibility cannot be excluded, but it leads to a view on which consciousness is itself irreducible, and is therefore to be classed in a nonreductive category (type D or type F).

There is one appeal to a "complete physics" that should be taken seriously. This is the idea that current physics characterizes its underlying properties (such as mass and charge) in terms of abstract structures and relations, but it leaves open their intrinsic natures. On this view, a complete physical description of the world must also characterize the intrinsic properties that ground these structures and relations; and once such intrinsic properties are invoked, physics will go beyond structure and dynamics, in such a way that truths about consciousness may be entailed. The relevant intrinsic properties are unknown to us, but they are knowable in principle. This is an important position, but it is precisely the position discussed under type F, so I defer discussion of it until then.

Second: What can be inferred from this sort of description in terms of structure and dynamics? A low-level microphysical description can entail all sorts of surprising and interesting macroscopic properties, as with the emergence of chemistry from physics, of biology from chemistry, or more generally of complex emergent behaviors in complex systems theory. But in all these cases, the complex properties that are entailed are nevertheless structural and dynamic: they describe complex spatiotemporal structures and complex dynamic patterns of behavior over those structures. So these cases support the general principle that from structure and dynamics, one can infer only structure and dynamics.

A type-C materialist might suggest there are some truths that are not themselves structural-dynamical that are nevertheless implied by a structural-dynamical description. It might be argued, perhaps, that truths about *representation* or *belief* have this character. But as we saw earlier, it seems clear that any sense in which these truths are implied by a structural-dynamic description involves a tacitly functional sense of representation or of belief. This is what we would expect: if claims involving these can be seen (on conceptual grounds) to be true *in virtue* of a structural-dynamic descriptions holding, the notions involved must themselves be structural-dynamic, at some level.

One might hold that there is some intermediate notion X, such that truths about X hold in virtue of structural-dynamic descriptions, and truths about consciousness hold in virtue of X. But as in the case of type-A materialism, either X is functionally analyzable (in the broad sense), in which case the second step fails, or X is not functionally analyzable, in which case the first step fails. This is brought out clearly in the case of representation: for the notion of functional representation, the first step fails, and for the notion of phenomenal representation, the second step fails. So this sort of strategy can only work by equivocation.

Third: does explaining or deducing complex structure and dynamics suffice to explain or deduce consciousness? It seems clearly not, for the usual reasons. Mary could know from her black-and-white room all about the spatiotemporal structure and dynamics of the world at all levels, but this will not tell her what it is like to see red. For any complex macroscopic structural or dynamic description of a system, one can conceive of that description being instantiated without consciousness. And explaining structure and dynamics of a human system is only to solve the easy problems, while leaving the hard problems untouched. To resist this last step, an opponent would have to hold that explaining structure and dynamics *thereby* suffices to explain consciousness. The only remotely tenable way to do this would be to embrace type-A materialism, which we have set aside.

A type-C materialist might suggest that instead of leaning on dynamics (as a type-A materialist does), one could lean on structure. Here, spatiotemporal structure seems very unpromising: to explain a system's size, shape, position, motion, and so on is clearly not to explain consciousness. A final possibility is leaning on the structure present in conscious states themselves. Conscious states have structure: there is both internal structure within a single complex conscious state, and there are patterns of similarities and differences between conscious states. But this structure is a distinctively *phenomenal* structure, quite different in kind from the spatiotemporal and formal structure present in physics. The structure of a complex phenomenal state is not spatiotemporal structure (although it may involve the representation of spatiotemporal structure), and the similarities and differences between phenomenal states are not formal similarities and differences, but differences between specific phenomenal characters. This is reflected in the fact that one can conceive of any spatiotemporal structure and formal structure without any associated phenomenal structure; one can know about the first without knowing about the second; and so on. So the epistemic gap is as wide as ever.

The basic problem with any type-C materialist strategy is that epistemic implication from A to B requires some sort of *conceptual hook* by virtue of which the condition described in A can satisfy the conceptual

requirements for the truth of B. When a physical account implies truths about life, for example, it does so in virtue of implying information about the macroscopic functioning of physical systems, of the sort required for life: here, broadly functional notions provide the conceptual hook. But in the case of consciousness, no such conceptual hook is available, given the structural-dynamic character of physical concepts, and the quite different character of the concept of consciousness.

Ultimately, it seems that any type-C strategy is doomed for familiar reasons. Once we accept that the concept of consciousness is not itself a functional concept, and that physical descriptions of the world are structural-dynamic descriptions, there is simply no conceptual room for it to be implied by a physical description. So the only room left is to hold that consciousness is a broadly functional concept after all (accepting type-A materialism), hold that there is more in physics than structure and dynamics (accepting type-D dualism or type-F monism), or holding that the truth of materialism does not require an implication from physics to consciousness (accepting type-B materialism).[*] So in the end, there is no separate space for the type-C materialist.

*[[Of those mentioned above as apparently sympathetic with type-C materialism, I think McGinn is ultimately a type-F monist, Nagel is either a type-B materialist or a type-F monist, and Churchland is either a type-B materialist or a type-Q materialist (below).]]

8 Interlude

Are there any other options for the materialist? One further option is to reject the distinctions on which this taxonomy rests. For example, some philosophers, especially followers of Quine (1951), reject any distinction between conceptual truth and empirical truth, or between the a priori and the a posteriori, or between the contingent and the necessary. One who is sufficiently Quinean might therefore reject the distinction between type-A and type-B materialism, holding that talk of epistemic implication and/or modal entailment is ungrounded, but that materialism is true nevertheless. We might call such a view type-Q materialism. Still, even on this view, similar issues arise. Some Quineans hold that explaining the functions explain everything (Dennett may be an example); if so, all the problems of type-A materialism arise. Others hold that we can postulate identities between physical states and conscious states in virtue of the strong isomorphic connections between them in nature (Paul Churchland may be an example); if so, the problems of type-B materialism arise. Others may appeal to novel future sorts of explanation; if so, the problems of type-C materialism arise. So the Quinean approach cannot avoid the relevant problems.

Leaving this sort of view aside, it looks like the only remotely viable options for the materialist are type-A materialism and type-B materialism. I think that other views are either ultimately unstable, or collapse into one of these (or the three remaining options).[*] It seems to me that the costs of these views — denying the manifest explanandum in the first case, and embracing primitive identities or strong necessities in the second case — suggest very strongly that they are to be avoided unless there are no viable alternatives.

*[[One might ask about specific reductive views, such as representationalism (which identifies consciousness with certain representational states), and higher-order thought theory (which identifies consciousness with the objects of higher-order

thoughts). How these views are classified depends on how a given theorist regards the representational or higher-order states (e.g., functionally definable or not) and their connection to consciousness (e.g., conceptual or empirical). Among representationalists, I think that Dretske 1995 and Harman 1990 are type-A materialists, while Lycan 1996 and Tye 1995 are type-B materialists. Among higher-order thought theorists, Carruthers 2000 is clearly a type-B materialist, while Rosenthal 1997 is either type-A or type-B. One could also in principle hold nonmaterialist versions of each of these views.]]

So the residual question is whether there are viable alternatives. If consciousness is not necessitated by physical truths, then it must involve something ontologically novel in the world: to use Kripke's metaphor, after fixing all the physical truths, God had to do more work to fix all the truths about consciousness. That is, there must be ontologically fundamental features of the world over and above the features characterized by physical theory. We are used to the idea that some features of the world are fundamental: in physics, features such as spacetime, mass, and charge, are taken as fundamental and not further explained. If the arguments against materialism are correct, these features from physics do not exhaust the fundamental features of the world: we need to expand our catalog of the world's basic features.

There are two possibilities here. First, it could be that consciousness is itself a fundamental feature of the world, like spacetime and mass. In this case, we can say that phenomenal properties are fundamental. Second, it could be that consciousness is not itself fundamental, but is necessitated by some more primitive fundamental feature X that is not itself necessitated by physics. In this case, we might call X a *protophenomenal* property, and we can say that protophenomenal properties are fundamental. I will typically put things in terms of the first possibility for ease of discussion, but the discussion that follows applies equally to the second. Either way, consciousness involves something novel and fundamental in the world.

The question then arises: how do these novel fundamental properties relate to the already acknowledged fundamental properties of the world, namely those invoked in microphysics? In general, where there are fundamental properties, there are fundamental laws. So we can expect that there will be some sort of fundamental principles — psychophysical laws — connecting physical and phenomenal properties. Like the fundamental laws of relativity or quantum mechanics, these psychophysical laws will not be deducible from more basic principles, but instead will be taken as primitive.

But what is the character of these laws? An immediate worry is that the microphysical aspects of the world is often held to be causally closed, in that every microphysical state has a microphysical sufficient cause. How are fundamental phenomenal properties to be integrated with this causally closed network?

There seem to be three main options for the nonreductionist here. First, one could deny the causal closure of the microphysical, holding that there are causal gaps in microphysical dynamics that are filled by a causal role for distinct phenomenal properties: this is type-D dualism. Second, one could accept the causal closure of the microphysical and hold that phenomenal properties play no causal role with respect to the physical network: this is type-E dualism. Third, one could accept that the microphysical network is causally closed, but hold that phenomenal properties are nevertheless integrated with it and play a causal role, by virtue of constituting the intrinsic nature of the physical: this is type-F monism.

In what follows, I will discuss each of these views. The discussion is necessarily speculative in certain respects, and I do not claim to establish that any one of the views is true or completely unproblematic. But I do aim to suggest that none of them has obvious fatal flaws, and that each deserves further investigation.

9 Type-D Dualism

Type-D dualism holds that microphysics is not causally closed, and that phenomenal properties play a causal role in affecting the physical world.[*] On this view, usually known as *interactionism*, physical states will cause phenomenal states, and phenomenal states cause physical states. The corresponding psychophysical laws will run in both directions. On this view, the evolution of microphysical states will not be determined by physical principles alone. Psychophysical principles specifying the effect of phenomenal states on physical states will also play an irreducible role.

*[[Type-D dualists include Foster 1991, Hodgson 1991, Popper and Eccles 1977, Sellars 1981, Stapp 1993, and Swinburne 1986.]]

The most familiar version of this sort of view is Descartes' substance dualism (hence D for Descartes), on which there are separate interacting mental and physical substances or entities. But this sort of view is also compatible with a property dualism, on which there is just one sort of substance or entity with both physical and phenomenal fundamental properties, such that the phenomenal properties play an irreducible role in affecting the physical properties. In particular, the view is compatible with an "emergentist" view such as Broad's, on which phenomenal properties are ontologically novel properties of physical systems (not deducible from microphysical properties alone), and have novel effects on microphysical properties (not deducible from microphysical principles alone). Such a view would involve basic principles of "downward" causation of the mental on the microphysical (hence also D for downward causation).

It is sometimes objected that distinct physical and mental states could not interact, since there is no causal nexus between them. But one lesson from Hume and from modern science is that the same goes for any fundamental causal interactions, including those found in physics. Newtonian science reveals no causal nexus by which gravitation works, for example; rather, the relevant laws are simply fundamental. The same goes for basic laws in other physical theories. And the same, presumably, applies to fundamental psychophysical laws: there is no need for a causal nexus distinct from the physical and mental properties themselves.

By far the most influential objection to interactionism is that it is incompatible with physics. It is widely held that science tells us that the microphysical realm is causally closed, so that there is no room for mental states to have any effects. An interactionist might respond in various ways. For example, it could be suggested that although no experimental studies have revealed these effects, none have ruled them out. It might further be suggested that physical theory allows any number of basic *forces* (four as things stand, but there is always room for more), and that an extra force associated with a mental field would be a reasonable extension of existing physical theory. These suggestions would invoke significant revisions to physical theory, so are not to be made lightly; but one could argue that nothing rules them out.

By far the strongest response to this objection, however, is to suggest that far from ruling out interactionism, contemporary physics is positively encouraging to the possibility. On the standard formulation of quantum mechanics, the state of the world is described by a wave function, according to which physical entities are often in a superposed state (e.g., in a superposition of two different positions), even though superpositions are never directly observed. On the standard dynamics, the wave function can evolve in two ways: linear evolution by the Schrödinger equation (which tends to produce superposed states), and nonlinear *collapses* from superposed states into nonsuperposed states. Schrödinger evolution is deterministic, but collapse is nondeterministic. Schrödinger evolution is constantly ongoing, but on the standard formulation, collapses occur only occasionally, on measurement.

The collapse dynamics leaves a door wide open for an interactionist interpretation. Any physical nondeterminism might be held to leave room for nonphysical effects, but the principles of collapse do much more than that. Collapse is supposed to occur on measurement. There is no widely agreed definition of what a measurement is, but there is one sort of event that everyone agrees is a measurement: observation by a conscious observer. Further, it seems that no purely physical criterion for a measurement can work, since purely physical systems are governed by the linear Schrödinger dynamics. As such, it is natural to suggest that a measurement is precisely a conscious observation, and that this conscious observation causes a collapse.

The claim should not be too strong: quantum mechanics does not force this interpretation of the situation onto us, and there are alternative interpretations of quantum mechanics on which there are no collapses, or on which measurement has no special role in collapse.[*] Nevertheless, quantum mechanics appears to be perfectly *compatible* with such an interpretation. In fact, one might argue that if one was to design elegant laws of physics that allow a role for the conscious mind, one could not do much better than the bipartite dynamics of standard quantum mechanics: one principle governing deterministic evolution in normal cases, and one principle governing nondeterministic evolution in special situations that have a *prima facie* link to the mental.

*[[No-collapse interpretations include Bohm's "hidden-variable" interpretations, and Everett's "many-worlds" (or "many-minds") interpretation. A collapse interpretation that does not invoke measurement is the Ghirardi-Rimini-Weber interpretation (with random occasional collapses). Each of these interpretations requires a significant revision to the standard dynamics of quantum mechanics, and each is controversial, although each has its benefits. (See Albert 1993 for discussion of these and other interpretations.) It is notable that there seems to be no remotely tenable interpretation that preserves the standard claim that collapses occur upon measurement, except for the interpretation involving consciousness.]]

Of course such an interpretation of quantum mechanics is controversial. Many physicists reject it precisely because it is dualistic, giving a fundamental role to consciousness. This rejection is not surprising, but it carries no force when we have independent reason to hold that consciousness may be fundamental. There is some irony in the fact that philosophers reject interactionism on largely physical grounds[*] (it is incompatible with physical theory), while physicists reject an interactionist interpretation of quantum mechanics on largely philosophical grounds (it is dualistic). Taken conjointly, these reasons carry little force, especially in light of the arguments against materialism elsewhere in this paper.

*[[I have been as guilty of this as anyone, setting aside interactionism in Chalmers 1996 partly for reasons of compatibility with physics. I am still not especially inclined to endorse interactionism, but I now think that the argument from physics is much too glib. Three further reasons for rejecting the view are mentioned in Chalmers 1996. First, if consciousness is to make an interesting qualitative difference to behavior, this requires that it act nonrandomly, in violation of the probabilistic requirements of quantum mechanics. I think there is something to this, but one could bite the bullet on nonrandomness in response, or one could hold that even a random causal role for consciousness is good enough. Second, I argued that denying causal closure yields no special advantage, as a view with causal closure can achieve much the same effect via type-F monism. Again there is something to this, but the type-D view does have the significant advantage of avoiding the type-F view's "combination problem." Third, it is not clear that the collapse interpretation yields the *sort* of causal role for consciousness that we expect it to have. I think that this is an important open question that requires detailed investigation.]]

This sort of interpretation needs to be formulated in detail to be assessed.[*] I think the most promising version of such an interpretation allows conscious states to be correlated with the total quantum state of a system, with the extra constraint that conscious states (unlike physical states) can never be superposed. In a conscious physical system such as a brain, the physical and phenomenal states of the system will be correlated in a (nonsuperposed) quantum state. Upon observation of a superposed system, then Schrödinger evolution at the moment of observation would cause the observed system to become correlated with the brain, yielding a resulting superposition of brain states and so (by psychophysical correlation) a superposition of conscious states. But such a superposition cannot occur, so one of the potential resulting conscious states is somehow selected (presumably by a nondeterministic dynamic principle at the phenomenal level). The result is that (by psychophysical correlation) a definite brain state and state of the observed object are also selected. The same might apply to the connection between consciousness and non-conscious processes in the brain: when superposed non-conscious processes threaten to affect consciousness, there will be some sort of selection. In this way, there is a causal role for consciousness in the physical world.

*[[Consciousness-collapse interpretations of quantum mechanics have been put forward by Wigner (1961), Hodgson (1991), and Stapp (1993). Only Stapp goes into much detail, with an interesting but somewhat idiosyncratic account that goes in a direction different from that suggested above.]]

(Interestingly, such a theory may be empirically testable. In quantum mechanics, collapse theories yield predictions slightly different from no-collapse theories, and different hypotheses about the location of collapse yield predictions that differ from each other, although the differences are extremely subtle and are currently impossible to measure. If the relevant experiments can one day be performed, some outcomes would give us strong reason to accept a collapse theory, and might in turn give us grounds to accept a role for consciousness. As a bonus, this could even yield an empirical criterion for the presence of consciousness.)

There are any number of further questions concerning the precise formulation of such a view, its compatibility with physical theory more generally (e.g., relativity and quantum field theory), and its philosophical tenability (e.g., does this view yield the sort of causal role that we are inclined to think consciousness must have). But at the very least, it cannot be said that physical theory immediately rules out the possibility of an interactionist theory. Those who make this claim often raise their eyebrows when

a specific theory such as quantum mechanics is mentioned; but this is quite clearly an inconsistent set of attitudes. If physics is supposed to rule out interactionism, then careful attention to the detail of physical theory is required.

All this suggests that there is at least room for a viable interactionism to be explored, and that the most common objection to interactionism has little force. Of course it does not entail that interactionism is true. There is much that is attractive about the view of the physical world as causally closed, and there is little direct evidence from cognitive science of the hypothesis that behavior cannot be wholly explained in terms of physical causes. Still, if we have independent reason to think that consciousness is irreducible, and if we wish to retain the intuitive view that consciousness plays a causal role, then this is a view to be taken very seriously.

10 Type-E Dualism

Type-E dualism holds that phenomenal properties are ontologically distinct from physical properties, and that the phenomenal has no effect on the physical.[*] This is the view usually known as *epiphenomenalism* (hence type-E): physical states cause phenomenal states, but not vice versa. On this view, psychophysical laws run in one direction only, from physical to phenomenal. The view is naturally combined with the view that the physical realm is causally closed: this further claim is not essential to type-E dualism, but it provides much of the motivation for the view.

*[[Type-E dualists include Campbell 1970, Huxley 1974, Jackson 1982, and Robinson 1988.]]

As with type-D dualism, type-E dualism is compatible with a substance dualism with distinct physical and mental substances or entities, and is also compatible with a property dualism with one sort of substance or entity and two sorts of properties. Again, it is compatible with an emergentism such as Broad's, on which mental properties are ontologically novel emergent properties of an underlying entity, but in this case although there are emergent qualities, there is no emergent downward causation.

Type-E dualism is usually put forward as respecting both consciousness and science: it simultaneously accommodates the anti-materialist arguments about consciousness and the causal closure of the physical. At the same time, type-E dualism is frequently rejected as deeply counterintuitive. If type-E dualism is correct, then phenomenal states have no effect on our actions, physically construed. For example, a sensation of pain will play no causal role in my hand's moving away from a flame; my experience of decision will play no causal role in my moving to a new country; and a sensation of red will play no causal role in my producing the utterance 'I am experiencing red now.' These consequences are often held to be obviously false, or at least unacceptable.

Still, the type-E dualist can reply that there is no direct *evidence* that contradicts their view. Our evidence reveals only regular connections between phenomenal states and actions, so that certain sorts of experiences are typically followed by certain sorts of actions. Being exposed to this sort of constant conjunction produces a strong *belief* in a causal connection (as Hume pointed out in another context); but

it is nevertheless compatible with the absence of a causal connection. Indeed, it seems that if epiphenomenalism *were* true, we would have exactly the same evidence, and be led to believe that consciousness has a causal role for much the same reasons. So if epiphenomenalism is otherwise coherent and acceptable, it seems that these considerations do not provide strong reasons to reject it.[*]

*[[Some accuse the epiphenomenalist of a double standard: relying on intuition in making the case against materialism, but going counter to intuition in denying a causal role for consciousness. But intuitions must be assessed against the background of reasons and evidence. To deny the relevant intuitions in the anti-materialist argument (in particular, the intuition of a further explanandum) appears to contradict the available first-person evidence; but denying a causal role for consciousness appears to be compatible on reflection with all our evidence, including first-person evidence.]]

Another objection holds that if consciousness is epiphenomenal, it could not have evolved by natural selection. The type-E dualist has a straightforward reply, however. On the type-E view, there are fundamental psychophysical laws associating physical and phenomenal properties. If evolution selects appropriate physical properties (perhaps involving physical or informational configurations in the brain), then the psychophysical laws will ensure that phenomenal properties are instantiated, too. If the laws have the right form, one can even expect that as more complex physical systems are selected, more complex states of consciousness will evolve. In this way, physical evolution will carry the evolution of consciousness along with it as a sort of byproduct.

Perhaps the most interesting objections to epiphenomenalism focus on the relation between consciousness and representations of consciousness. It is certainly at least strange to suggest that consciousness plays no causal role in my utterances of 'I am conscious'. Some have suggested more strongly that this rules out any *knowledge* of consciousness. It is often held that if a belief about X is to qualify as knowledge, the belief must be caused in some fashion by X. But if consciousness does not effect physical states, and if beliefs are physically constituted, then consciousness cannot cause beliefs. And even if beliefs are not physically constituted, it is not clear how epiphenomenalism can accommodate a causal connection between consciousness and belief.

In response, an epiphenomenalist can deny that knowledge always requires a causal connection. One can argue on independent grounds that there is a stronger connection between consciousness and beliefs about consciousness: consciousness plays a role in *constituting* phenomenal concepts and phenomenal beliefs. A red experience plays a role in constituting a belief that one is having a red experience, for example. If so, there is no causal distance between the experience and the belief. And one can argue that this immediate connection to experience and belief allows for the belief to be justified. If there is right, then epiphenomenalism poses no obstacle to knowledge of consciousness.

A related objection holds that my zombie twin would produce the same reports (e.g., 'I am conscious'), caused by the same mechanisms, and that his reports are unjustified; if so, my own reports are unjustified. In response, one can hold that the true bearers of justification are beliefs, and that my zombie twin and I have *different* beliefs, involving different concepts, because of the role that consciousness plays in constituting my concepts but not the zombie's. Further, the fact that we produce isomorphic reports implies that a third-person observer might not be any more justified in believing that I am conscious than

that the zombie is conscious, but it does not imply a difference in first-person justification. The first-person justification for my belief that I am conscious is not grounded in any way in my reports but rather in my experiences themselves, experiences that the zombie lacks.

I think that there is no knockdown objection to epiphenomenalism here. Still, it must be acknowledged that the situation is at least odd and counterintuitive. The oddness of epiphenomenalism is exacerbated by the fact that the relationship between consciousness and reports about consciousness seems to be something of a lucky coincidence, on the epiphenomenalist view. After all, if psychophysical laws are independent of physical evolution, then there will be possible worlds where physical evolution is the same as ours but the psychophysical laws are very different, so that there is a radical mismatch between reports and experiences. It seems lucky that we are in a world whose psychophysical laws match them up so well. In response, an epiphenomenalist might try to make the case that these laws are somehow the most "natural" and are to be expected; but there is at least a significant burden of proof here.

Overall, I think that epiphenomenalism is a coherent view without fatal problems. At the same time, it is an inelegant view, producing a fragmented picture of nature, on which physical and phenomenal properties are only very weakly integrated in the natural world. And of course it is a counterintuitive view that many people find difficult to accept. Inelegance and counterintuitiveness are better than incoherence; so if good arguments force us to epiphenomenalism as the most coherent view, then we should take it seriously. But at the same time, we have good reason to examine other views very carefully.

11 Type-F Monism

Type-F monism is the view that consciousness is constituted by the intrinsic properties of fundamental physical entities: that is, by the categorical bases of fundamental physical dispositions.[*] On this view, phenomenal or protophenomenal properties are located at the fundamental level of physical reality, and in a certain sense, underlie physical reality itself.

*[[Versions of type-F monism have been put forward by Russell 1926, Feigl 1958/1967, Maxwell 1979, Lockwood 1989, Chalmers 1996, Griffin 1998, Strawson 2000, and Stoljar 2001.]]

This view takes its cue from Bertrand Russell's discussion of physics in *The Analysis of Matter*. Russell pointed out that physics characterizes physical entities and properties by their relations to one another and to us. For example, a quark is characterized by its relations to other physical entities, and a property such as mass is characterized by an associated dispositional role, such as the tendency to resist acceleration. At the same time, physics says nothing about the intrinsic nature of these entities and properties. Where we have relations and dispositions, we expect some underlying intrinsic properties that ground the dispositions, characterizing the entities that stand in these relations.[*] But physics is silent about the intrinsic nature of a quark, or about the intrinsic properties that play the role associated with mass. So this is one metaphysical problem: what are the intrinsic properties of fundamental physical systems?

*[[There is philosophical debate over the thesis that all dispositions have a categorical basis. If the thesis is accepted, the

case for type-F monism is particularly strong, since microphysical dispositional must have a categorical basis, and we have no independent characterization of that basis. But even if the thesis is rejected, type-F monism is still viable. We need only the thesis that microphysical dispositions *may* have a categorical basis to open room for intrinsic properties here.]]

At the same time, there is another metaphysical problem: how can phenomenal properties be integrated with the physical world? Phenomenal properties seem to be intrinsic properties that are hard to fit in with the structural/dynamic character of physical theory; and arguably, they are the only intrinsic properties that we have direct knowledge of. Russell's insight was that we might solve both these problems at once. Perhaps the intrinsic properties of the physical world are themselves phenomenal properties. Or perhaps the intrinsic properties of the physical world are not phenomenal properties, but nevertheless constitute phenomenal properties: that is, perhaps they are protophenomenal properties. If so, then consciousness and physical reality are deeply intertwined.

This view holds the promise of integrating phenomenal and physical properties very tightly in the natural world. Here, nature consists of entities with intrinsic (proto)phenomenal qualities standing in causal relations within a spacetime manifold. Physics as we know it emerges from the relations between these entities, whereas consciousness as we know it emerges from their intrinsic nature. As a bonus, this view is perfectly compatible with the causal closure of the microphysical, and indeed with existing physical laws. The view can retain the *structure* of physical theory as it already exists; it simply supplements this structure with an intrinsic nature. And the view acknowledges a clear causal role for consciousness in the physical world: (proto)phenomenal properties serve as the ultimate categorical basis of all physical causation.

This view has elements in common with both materialism and dualism. From one perspective, it can be seen as a sort of materialism. If one holds that physical terms refer not to dispositional properties but the underlying intrinsic properties, then the protophenomenal properties can be seen as physical properties, thus preserving a sort of materialism. From another perspective, it can be seen as a sort of dualism. The view acknowledges phenomenal or protophenomenal properties as ontologically fundamental, and it retains an underlying duality between structural-dispositional properties (those directly characterized in physical theory) and intrinsic protophenomenal properties (those responsible for consciousness). One might suggest that while the view arguably fits the letter of materialism, it shares the spirit of antimaterialism.

In its protophenomenal form, the view can be seen as a sort of neutral monism: there are underlying neutral properties X (the protophenomenal properties), such that the X properties are simultaneously responsible for constituting the physical domain (by their relations) and the phenomenal domain (by their collective intrinsic nature). In its phenomenal form, can be seen as a sort of idealism, such that mental properties constitute physical properties, although these need not be mental properties in the mind of an observer, and they may need to be supplemented by causal and spatiotemporal properties in addition. One could also characterize this form of the view as a sort of panpsychism, with phenomenal properties ubiquitous at the fundamental level. One could give the view in its most general form the name *panprotopsyichism*, with either protophenomenal or phenomenal properties underlying all of physical reality.

A type-F monist may have one of a number of attitudes to the zombie argument against materialism. Some type-F monists may hold that a complete physical description must be expanded to include an intrinsic description, and may consequently deny that zombies are conceivable. (We only think we are conceiving of a physically identical system because we overlook intrinsic properties.) Others could maintain that existing physical concepts refer via dispositions to those intrinsic properties that ground the dispositions. If so, these concepts have different primary and secondary intensions, and a type-F monist could correspondingly accept conceivability but deny possibility: we misdescribe the conceived world as physically identical to ours, when in fact it is just structurally identical.[*] Finally, a type-F monist might hold that physical concepts refer to dispositional properties, so that zombies are both conceivable and possible, and the intrinsic properties are not physical properties. The differences between these three attitudes seem to be ultimately terminological rather than substantive.

*[[Hence type-F monism is the sort of "physicalism" that emerges from the loophole mentioned in the two-dimensional argument against type-B materialism. The only way a "zombie world" W could satisfy the primary intension but not the secondary intension of P is for it to share the dispositional structure of our world but not the underlying intrinsic microphysical properties. If this difference is responsible for the lack of consciousness in W, then the intrinsic microphysical properties in our world are responsible for constituting consciousness. Maxwell (1979) exploits this sort of loophole in replying to Kripke's argument.

Note that such a W must involve either a different corpus of intrinsic properties from those in our world, or no intrinsic properties at all. A type-F monist who holds that the only coherent intrinsic properties are protophenomenal properties might end up denying the conceivability of zombies, even under a structural-functional description of their physical state — for reasons very different from those of the type-A materialist.]]

As for the knowledge argument, a type-F monist might insist that for Mary to have complete physical knowledge, she would have to have a description of the world involving concepts that directly characterize the intrinsic properties; if she had this (as opposed to her impoverished description involving dispositional concepts), she might thereby be in a position to know what it is like to see red. Regarding the explanatory argument, a type-F monist might hold that physical accounts involving intrinsic properties can explain more than structure and function. Alternatively, a type-F monist who sticks to dispositional physical concepts will make responses analogous to one of the other two responses above.

The type-F view is admittedly speculative, and it can sound strange at first hearing. Many find it extremely counterintuitive to suppose that fundamental physical systems have phenomenal properties: e.g., that there is something it is like to be an electron. The protophenomenal version of the view rejects this claim, but retains something of its strangeness: it seems that any properties responsible for constituting consciousness must be strange and unusual properties, of a sort that we might not expect to find in microphysical reality. Still, it is not clear that this strangeness yields any strong objections. Like epiphenomenalism, the view appears to be compatible with all our evidence, and there is no direct evidence against it. One can argue that if the view were true, things would appear to us just as they in fact appear. And we have learned from modern physics that the world is a strange place: we cannot expect it to obey all the dictates of common sense.

One might also object that we do not have any conception of what protophenomenal properties might be like, or of how they could constitute phenomenal properties. This is true, but one could suggest that this is merely a product of our ignorance. In the case of familiar physical properties, there were principled reasons (based on the character of physical concepts) for denying a constitutive connection to phenomenal properties. Here, there are no such principled reasons. At most, there is ignorance and absence of constitution. Of course it would be very desirable to form a positive conception of protophenomenal properties. Perhaps we can do this indirectly, by some sort of theoretical inference from the character of phenomenal properties to their underlying constituents; or perhaps knowledge of the nature of protophenomenal properties will remain beyond us. Either way, this is no reason to reject the truth of the view.[*]

*[[McGinn (1991) can be read as advocating a type-F view, while denying that we can know the nature of the protophenomenal properties. His arguments rest on the claim that these properties cannot be known either through perception or through introspection. But this does not rule out the possibility that they might be known through some sort of inference to the best explanation of (introspected) phenomenology, subject to the additional constraints of (perceived) physical structure.]]

There is one sort of principled problem in the vicinity. Our phenomenology has a rich and specific structure: it is unified, bounded, differentiated into many different aspects, but with an underlying homogeneity to many of the aspects, and appears to have a single subject of experience. It is not easy to see how a distribution of a large number of individual microphysical systems, each with their own protophenomenal properties, could somehow add up to this rich and specific structure. Should one not expect something more like a disunified, jagged collection of phenomenal spikes?

This is a version of what James called the *combination problem* for panpsychism, or what Stoljar (2001) calls the *structural mismatch* problem for the Russellian view (see also Foster 1991, pp. 119-30). To answer it, it seems that we need a much better understanding of the *compositional* principles of phenomenology: that is, the principles by which phenomenal properties can be composed or constituted from underlying phenomenal properties, or protophenomenal properties. We have a good understanding of the principles of physical composition, but no real understanding of the principles of phenomenal composition. This is an area that deserves much close attention: I think it is easily the most serious problem for the type-F monist view. At this point, it is an open question whether or not the problem can be solved.

Some type-F monists appear to hold that they can avoid the combination problem by holding that phenomenal properties are the intrinsic properties of *high-level* physical dispositions (e.g., those involved in neural states), and need not be constituted by the intrinsic properties of microphysical states (hence they may also deny panprotopsychoism). But this seems to be untenable: if the low-level network is causally closed and the high-level intrinsic properties are not constituted by low-level intrinsic properties, the high-level intrinsic properties will be epiphenomenal all over again, for familiar reasons. The only way to embrace this position would seem to be in combination with a denial of microphysical causal closure, holding that there are fundamental dispositions above the microphysical level, which have phenomenal properties as their grounds. But such a view would be indistinguishable from type-D

dualism.[*] So a distinctive type-F monism will have to face the combination problem directly.

*[[In this way, we can see that type-D views and type-F views are quite closely related. We can imagine that if a type-D view is true and there are microphysical causal gaps, we could be led through physical observation alone to postulate higher-level entities to fill these gaps — "psychons", say --where these are characterized in wholly structural/dispositional terms. The type-D view adds to this the suggestion that psychons have an intrinsic phenomenal nature. The main difference between the type-D view and the type-F view is that the type-D view involves fundamental causation above the microphysical level. This will involve a more radical view of physics, but it might have the advantage of avoiding the combination problem.]]

Overall, type-F monism promises a deeply integrated and elegant view of nature. No-one has yet developed any sort of detailed theory in this class, and it is not yet clear whether such a theory can be developed. But at the same time, there appear to be no strong reasons to reject the view. As such, type-F monism is likely to provide fertile grounds for further investigation, and it may ultimately provide the best integration of the physical and the phenomenal within the natural world.

12 Conclusions

Are there any other options for the nonreductionist? There are two views that may not fit straightforwardly into the categories above.

First, some nonmaterialists hold that phenomenal properties are ontologically wholly distinct from physical properties, that microphysics is causally closed, but that phenomenal properties play a causal role with respect to the physical nevertheless. One way this might happen is by a sort of causal overdetermination: physical states causally determine behavior, but phenomenal states cause behavior at the same time. Another is by causal mediation: it might be that in at least some instances of microphysical causation from A to B, there is actually a causal connection from A to the mind to B, so that the mind enters the causal nexus without altering the structure of the network. And there may be further strategies here. We might call this class type-O dualism (taking overdetermination as a paradigm case). These views shares much of the structure of the type-E view (causally closed physical world, distinct phenomenal properties), but escapes the charge of epiphenomenalism. The special causal setups of these views may be hard to swallow, and they share some of the same problems as the type-E view (e.g., the fragmented view of nature, and the "lucky" psychophysical laws), but this class should nevertheless be put on the table as an option.[*]

*[[Type-O positions are advocated by Bealer (forthcoming), Lowe 1996 and Mills 1996.]]

Second, some nonmaterialists are *idealists* (in a Berkeleyan sense), holding that the physical world is itself constituted by the conscious states of an observing agent. We might call this view type-I monism. It shares with type-F monism the property that phenomenal states play a role in constituting physical reality, but on the type-I view this happens in a very different way: not by having separate "microscopic" phenomenal states underlying each physical state, but rather by having physical states constituted holistically by a "macroscopic" phenomenal mind. This view seems to be non-naturalistic in a much

deeper sense than any of the views above, and in particular seems to suffer from an absence of causal or explanatory closure in nature: once the natural explanation in terms of the external world is removed, highly complex regularities among phenomenal states have to be taken as unexplained in terms of simpler principles. But again, this sort of view should at least be acknowledged.

As I see things, the best options for a nonreductionist are type-D dualism, type-E dualism, or type-F monism: that is, interactionism, epiphenomenalism, or panprotopsyism. If we acknowledge the epistemic gap between the physical and the phenomenal, and we rule out primitive identities and strong necessities, then we are led to a disjunction of these three views. Each of the views has at least some promise, and none have clear fatal flaws. For my part, I give some credence to each of them. I think that in some ways the type-F view is the most appealing, but this sense is largely grounded in aesthetic considerations whose force is unclear.

The choice between these three views may depend in large part on the development of specific theories within these frameworks. Especially for the type-D view and type-F view, further theoretical work is crucial in assessing the theories (e.g., in explicating quantum interactionism, or in understanding phenomenal composition). It may also be that the empirical science of consciousness will give some guidance. As the science progress, we will be led to infer simple principles that underlie correlations between physical and phenomenal states. It may be that these principles turn out to point strongly toward one or the other of these views: e.g., if simple principles connecting microphysical states to phenomenal or protophenomenal states can do the explanatory work, then we may have reason to favor a type-F view, while if the principles latch onto the physical world at a higher level, then we may have reason to favor a type-D or type-E view. And if consciousness has a specific pattern of effects on the physical world, as the type-D view suggests, then empirical studies ought in principle to be able to find these effects, although perhaps only with great difficulty.

Not everyone will agree that each of these views is viable. It may be that further examination will reveal deep problems with some of these views. But this further examination needs to be performed. There has been little critical examination of type-F views to date, for example; we have seen that the standard arguments against type-D views carry very little weight; and while arguments against type-E views carry some intuitive force, they are far from making a knockdown case against the views. I suspect that even if further examination reveals deep problems for some views in this vicinity, it is very unlikely that all such views will be eliminated.

In any case, this gives us some perspective on the mind-body problem. It is often held that even though it is hard to see how materialism could be true, materialism *must* be true, since the alternatives are unacceptable. As I see it, there are at least three *prima facie* acceptable alternatives to materialism on the table, each of which is compatible with a broadly naturalistic (even if not materialistic) worldview, and none of which has fatal problems. So given the clear arguments against materialism, it seems to me that we should at least tentatively embrace the conclusion that one of these views is correct. Of course all of the views discussed in this paper need to be developed in much more detail, and examined in light of all relevant scientific and philosophical developments, in order to be comprehensively assessed. But as things stand, I think that we have good reason to suppose that consciousness has a fundamental place in nature.

References

- Albert, D. Z. 1993. *Quantum Mechanics and Experience*. Harvard University Press.
- Bealer, G. 1994. Mental properties. *Journal of Philosophy* 91:185-208.
- Bealer, G. (forthcoming). Mental causation.
- Block, N. and Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review* 108:1-46.
- Broad, C. D. 1925. *The Mind and its Place in Nature*. Routledge and Kegan Paul.
- Campbell, K. K. 1970. *Body and Mind*. Doubleday.
- Carruthers, P. 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.
- Chalmers, D. J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:200-19. Reprinted in Shear 1997. <http://consc.net/papers/facing.html>.
- Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D. J. 1997. Moving forward on the problem of consciousness. *Journal of Consciousness Studies* 4:3-46. Reprinted in Shear 1997. <http://consc.net/papers/moving.html>.
- Chalmers, D. J. 1999. Materialism and the metaphysics of modality. *Philosophy and Phenomenological Research* 59:473-93. <http://consc.net/papers/modality.html>.
- Chalmers, D. J. 2002a. The content and epistemology of phenomenal belief. In Q. Smith and A. Jokic (eds.), *Aspects of Consciousness*. Oxford University Press. <http://consc.net/papers/belief.html>.
- Chalmers, D. J. 2002b. Does conceivability entail possibility? In T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford University Press. <http://consc.net/papers/conceivability.html>.
- Chalmers, D. J. forthcoming. The foundations of two-dimensional semantics. <http://consc.net/papers/foundations.html>.
- Chalmers, D. J. and Jackson, F. 2001. Conceptual analysis and reductive explanation. *Philosophical Review* 110:315-61. <http://consc.net/papers/analysis.html>.

- Churchland, P. M. 1996. The rediscovery of light. *Journal of Philosophy* 93:211-28.
- Churchland, P. S. 1997. The hornswoggle problem. In Shear 1997.
- Clark, A. 2000. A case where access implies qualia? *Analysis* 60:30-38.
- Dennett, D. C. 1991. *Consciousness Explained*. Little-Brown.
- Dennett, D. C. 1996. Facing backward on the problem of consciousness. *Journal of Consciousness Studies* 3:4-6.
- Dennett, D. C. 2001. The fantasy of first-person science. Forthcoming.
<http://ase.tufts.edu/cogstud/papers/chalmersdeb3dft.htm>.
- Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.
- Evans, G. 1979. Reference and contingency. *The Monist* 62:161-89.
- Feigl, H. 1958/1967. The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science* 2:370-497. Reprinted (with a postscript) as *The 'Mental' and the 'Physical'*. University of Minnesota Press.
- Foster, J. 1991. *The Immaterial Self: A Defence of the Cartesian Dualist Conception of the Mind*. Oxford University Press.
- Griffin, D. R. 1998. *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem*. University of California Press.
- Harman, G. 1990. The intrinsic quality of experience. *Philosophical Perspectives* 4:31-52.
- Hill, C. S. 1997. Imaginability, conceivability, possibility, and the mind-body problem. *Philosophical Studies* 87:61-85.
- Hodgson, D. 1991. *The Mind Matters: Consciousness and Choice in a Quantum World*. Oxford University Press.
- Huxley, T. 1874. On the hypothesis that animals are automata, and its history. *Fortnightly Review* 95:555-80. Reprinted in *Collected Essays*. London, 1893.
- Jackson, F. 1979. A note on physicalism and heat. *Australasian Journal of Philosophy* 58:26-34.

- Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136.
- Jackson, F. 1994. Finding the mind in the natural world. In R. Casati, B. Smith, and G. White (eds.), *Philosophy and the Cognitive Sciences*. Vienna: Holder-Pichler-Tempsky.
- Kaplan, D. 1989. Demonstratives. In J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*. New York: Oxford University Press.
- Kirk, R. 1974. Zombies vs materialists. *Proceedings of the Aristotelian Society (Supplementary Volume)* 48:135-52.
- Kripke, S. A. 1980. *Naming and Necessity*. Harvard University Press.
- Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.
- Levine, J. 2000. *Purple Haze: The Puzzle of Conscious Experience*. MIT Press.
- Lewis, D. 1988. What experience teaches. *Proceedings of the Russellian Society* (University of Sydney).
- Lewis, D. 1994. Reduction of mind. In S. Guttenplan (ed.), *Companion to the Philosophy of Mind*. Blackwell.
- Loar, B. 1990/1997. Phenomenal states. *Philosophical Perspectives* 4:81-108. Revised edition in (N. Block, O. Flanagan, and G. Güzeldere, eds) *The Nature of Consciousness*. MIT Press.
- Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford University Press.
- Lockwood, M. 1993. The grain problem. In H. Robinson (ed.), *Objections to Physicalism*. Oxford University Press.
- Lowe, E.J. 1996. *Subjects of Experience*. Cambridge University Press.
- Lycan, W.G. 1996. *Consciousness and Experience*. MIT Press.
- Maxwell, N. 1968. Understanding sensations. *Australasian Journal of Philosophy* 46:127-45.
- Maxwell, G. 1979. Rigid designators and mind-brain identity. *Minnesota Studies in the Philosophy of Science* 9:365-403.
- McGinn, C. 1989. Can we solve the mind-body problem? *Mind* 98:349-66.

- Mills, E. 1996. Interactionism and overdetermination. *American Philosophical Quarterly* 33:105-115.
- Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 83:435-50.
- Nordby, K. 1990. Vision in a complete achromat: A personal account. In R. Hess, L. Sharpe, and K. Nordby (eds.), *Night Vision: Basic, Clinical, and Applied Aspects*. Cambridge University Press.
- Papineau, D. 1993. Physicalism, consciousness, and the antipathetic fallacy. *Australasian Journal of Philosophy* 71:169-83.
- Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. MIT Press.
- Popper, K. and Eccles, J. 1977. *The Self and Its Brain: An Argument for Interactionism*. Springer.
- Quine, W. V. 1951. Two dogmas of empiricism. *Philosophical Review* 60:20-43.
- Rey, G. 1995. Toward a projectivist account of conscious experience. In T. Metzinger (ed.), *Conscious Experience*. Ferdinand Schoningh.
- Robinson, W. S. 1988. *Brains and People: An Essay on Mentality and its Causal Conditions*. Temple University Press.
- Rosenthal, D. M. 1997. A theory of consciousness. In N. Block, O. Flanagan, and G. Güzeldere (eds.), *The Nature of Consciousness*. MIT Press.
- Russell, B. 1927. *The Analysis of Matter*. London: Kegan Paul.
- Ryle, G. 1949. *The Concept of Mind*. Hutchinson and Co.
- Sellars, W. 1981. Is consciousness physical? *The Monist* 64:66-90.
- Shear, J. (ed.) 1997. *Explaining Consciousness: The Hard Problem*. MIT Press.
- Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:291-315.
- Smart, J. J. C. 1959. Sensations and brain processes. *Philosophical Review* 68:141-56.
- Stalnaker, R. 1978. Assertion. In P. Cole (ed.), *Syntax and Semantics: Pragmatics, Vol. 9*. New York: Academic Press.
- Stapp, H. 1993. *Mind, Matter, and Quantum Mechanics*. Springer-Verlag.

Stoljar, D. 2001. Two conceptions of the physical. *Philosophy and Phenomenological Research* 62:253-81.

Strawson, G. 2000. Realistic materialist monism. In S. Hameroff, A. Kaszniak, and D. Chalmers (eds.), *Toward a Science of Consciousness III*. MIT Press.

Swinburne, R. 1986. *The Evolution of the Soul*. Oxford University Press.

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press.

Van Gulick, R. 1993. Understanding the phenomenal mind: Are we all just armadillos? In M. Davies and G. Humphreys (eds.) *Consciousness: Philosophical and Psychological Aspects*. Blackwell.

White, S. 1986. Curse of the qualia. *Synthese* 68:333-68.

Wigner, E. P. 1961. Remarks on the mind-body question. In I. J. Good (ed.), *The Scientist Speculates*. Basic Books.

The Puzzle of Conscious Experience

[David J. Chalmers](#)

**Department of Philosophy
University of Arizona
Tucson, AZ 85721**

chalmers@arizona.edu

[*Scientific American*, December 1995 pp. 62-68. N.B. As always at *Scientific American*, this isn't just me talking. For a more careful treatment of this material, see my "[Facing Up to the Problem of Consciousness](#)".]

Conscious experience is at once the most familiar thing in the world and the most mysterious. There is nothing we know about more directly than consciousness, but it is extraordinarily hard to reconcile it with everything else we know. Why does it exist? What does it do? How could it possibly arise from neural processes in the brain? These questions are among the most intriguing in all of science.

From an objective viewpoint, the brain is relatively comprehensible. When you look at this page, there is a whirl of processing: photons strike your retina, electrical signals are passed up your optic nerve and between different areas of your brain, and eventually you might respond with a smile, a perplexed frown or a remark. But there is also a subjective aspect. When you look at the page, you are conscious of it, directly experiencing the images and words as part of your private, mental life. You have vivid impressions of colored flowers and vibrant sky. At the same time, you may be feeling some emotions and forming some thoughts. Together such experiences make up consciousness: the subjective, inner life of the mind.

For many years, consciousness was shunned by researchers studying the brain and the mind. The prevailing view was that science, which depends on objectivity, could not accommodate something as subjective as consciousness. The behaviorist movement in psychology, dominant earlier in this century concentrated on external behavior and disallowed any talk of internal mental processes. Later, the rise of cognitive science focused attention on processes inside the head. Still, consciousness remained off-limits, fit only for late-night discussion over drinks.

Over the past several years, however, an increasing number of neuroscientists, psychologists and philosophers have been rejecting the idea that consciousness cannot be studied and are attempting to delve into its secrets. As might be expected of a field so new, there is a tangle of diverse and conflicting theories, often using basic concepts in incompatible ways. To help unsnarl the tangle, philosophical reasoning is vital.

The myriad views within the field range from reductionist theories, according to which consciousness

can be explained by the standard methods of neuroscience and psychology, to the position of the so-called mysterians, who say we will never understand consciousness at all. I believe that on close analysis both of these views can be seen to be mistaken and that the truth lies somewhere in the middle. Against reductionism I will argue that the tools of neuroscience cannot provide a full account of conscious experience, although they have much to offer. Against mysterianism I will hold that consciousness might be explained by a new kind of theory. The full details of such a theory are still out of reach, but careful reasoning and some educated inferences can reveal something of its general nature. For example, it will probably involve new fundamental laws, and the concept of information may play a central role. These faint glimmerings suggest that a theory of consciousness may have startling consequences for our view of the universe and of ourselves.

The Hard Problem

Researchers use the word "consciousness" in many different ways. To clarify the issues, we first have to separate the problems that are often clustered together under the name. For this purpose, I find it useful to distinguish between the "easy problems" and the "hard problem" of consciousness. The easy problems are by no means trivial - they are actually as challenging as most in psychology and biology - but it is with the hard problem that the central mystery lies.

The easy problems of consciousness include the following: How can a human subject discriminate sensory stimuli and react to them appropriately? How does the brain integrate information from many different sources and use this information to control behavior? How is it that subjects can verbalize their internal states? Although all these questions are associated with consciousness, they all concern the objective mechanisms of the cognitive system. Consequently, we have every reason to expect that continued work in cognitive psychology and neuroscience will answer them.

The hard problem, in contrast, is the question of how physical processes in the brain give rise to subjective experience. This puzzle involves the inner aspect of thought and perception: the way things feel for the subject. When we see, for example, we experience visual sensations, such as that of vivid blue. Or think of the ineffable sound of a distant oboe, the agony of an intense pain, the sparkle of happiness or the meditative quality of a moment lost in thought. All are part of what I am calling consciousness. It is these phenomena that pose the real mystery of the mind.



ISOLATED NEUROSCIENTIST in a black-and-white room knows everything about how the brain processes colors but does not know what it is like to see them. This scenario suggests that knowledge of the brain

does not yield complete knowledge of conscious experience.

To illustrate the distinction, consider a thought experiment devised by the Australian philosopher Frank Jackson. Suppose that Mary, a neuroscientist in the 23rd century, is the world's leading expert on the brain processes responsible for color vision. But Mary has lived her whole life in a black-and-white room and has never seen any other colors. She knows everything there is to know about physical processes in the brain - its biology, structure and function. This understanding enables her to grasp everything there is to know about the easy problems: how the brain discriminates stimuli, integrates information and produces verbal reports. From her knowledge of color vision, she knows the way color names correspond with wavelengths on the light spectrum. But there is still something crucial about color vision that Man does not know: what it is like to experience a color such as red. It follows that there are facts about conscious experience that cannot be deduced from physical facts about the functioning of the brain.

Indeed, nobody knows why these physical processes are accompanied by conscious experience at all. Why is it that when our brains process light of a certain wavelength, we have an experience of deep purple? Why do we have any experience at all? Could not an unconscious automaton have performed the same tasks just as well? These are questions that we would like a theory of consciousness to answer.

I am not denying that consciousness arises from the brain. We know, for example, that the subjective experience of vision is closely linked to processes in the visual cortex. It is the link itself that perplexes, however. Remarkably, subjective experience seems to emerge from a physical process. But we have no idea how or why this is.

Is Neuroscience Enough?

Given the flurry of recent work on Consciousness in neuroscience and psychology, one might think this mystery is starting to be cleared up. On closer examination, however, it turns out that almost all the current work addresses only the easy problems of consciousness. The confidence of the reductionist view comes from the progress on the easy problems, but none of this makes any difference where the hard problem is concerned.

Consider the hypothesis put forward by neurobiologists Francis Crick of the Salk Institute for Biological Studies in San Diego and Christof Koch of the California Institute of Technology. They suggest that consciousness may arise from certain oscillations in the cerebral cortex, which become synchronized as neurons fire 40 times per second. Crick and Koch believe the phenomenon might explain how different attributes of a single perceived object (its color and shape, for example), which are processed in different parts of the brain, are merged into a coherent whole. In this theory, two pieces of information become bound together precisely when they are represented by synchronized neural firings.

The hypothesis could conceivably elucidate one of the easy problems about how information is integrated in the brain. But why should synchronized oscillations give rise to a visual experience, no matter how much integration is taking place? This question involves the hard problem, about which the

theory has nothing to offer. Indeed, Crick and Koch are agnostic about whether the hard problem can be solved by science at all.

The same kind of critique could be applied to almost all the recent work on consciousness. In his 1991 book *Consciousness Explained*, philosopher Daniel C. Dennett laid out a sophisticated theory of how numerous independent processes in the brain combine to produce a coherent response to a perceived event. The theory might do much to explain how we produce verbal reports on our internal states, but it tells us very little about why there should be a subjective experience behind these reports. Like other reductionist theories, Dennett's is a theory of the easy problems.

The critical common trait among these easy problems is that they all concern how a cognitive or behavioral function is performed. All are ultimately questions about how the brain carries out some task—how it discriminates stimuli, integrates information, produces reports and so on. Once neurobiology specifies appropriate neural mechanisms, showing how the functions are performed, the easy problems are solved. The hard problem of consciousness, in contrast, goes beyond problems about how functions are performed. Even if every behavioral and cognitive function related to consciousness were explained, there would still remain a further mystery: Why is the performance of these functions accompanied by conscious experience? It is this additional conundrum that makes the hard problem hard.

The Explanatory Gap

Some have suggested that to solve the hard problem, we need to bring in new tools of physical explanation: nonlinear dynamics, say, or new discoveries in neuroscience, or quantum mechanics. But these ideas suffer from exactly the same difficulty. Consider a proposal from Stuart R. Hameroff of the University of Arizona and Roger Penrose of the University of Oxford. They hold that consciousness arises from quantum-physical processes taking place in microtubules, which are protein structures inside neurons. It is possible (if not likely) that such a hypothesis will lead to an explanation of how the brain makes decisions or even how it proves mathematical theorems, as Hameroff and Penrose suggest. But even if it does, the theory is silent about how these processes might give rise to conscious experience. Indeed, the same problem arises with any theory of consciousness based only on physical processing.

The trouble is that physical theories are best suited to explaining why systems have a certain physical structure and how they perform various functions. Most problems in science have this form; to explain life, for example, we need to describe how a physical system can reproduce, adapt and metabolize. But consciousness is a different sort of problem entirely, as it goes beyond the explanation of structure and function.

Of course, neuroscience is not irrelevant to the study of consciousness. For one, it may be able to reveal the nature of the neural correlate of consciousness - the brain processes most directly associated with conscious experience. It may even give a detailed correspondence between specific processes in the brain and related components of experience. But until we know why these processes give rise to conscious experience at all, we will not have crossed what philosopher Joseph Levine has called the explanatory

gap between physical processes and consciousness. Making that leap will demand a new kind of theory.



BLOOD FLOW variations in the visual cortex demonstrate how a subject's brain responds to a pattern being viewed. The colors in this image show the cortical activity corresponding to the subject's view of either half of the pattern. This activity is a neural correlate of visual experience, but the relationship between the two remains mysterious.

A True Theory of Everything

In searching for an alternative, a key observation is that not all entities in science are explained in terms of more basic entities. In physics, for example, space-time, mass and charge (among other things) are regarded as fundamental features of the world, as they are not reducible to anything simpler. Despite this irreducibility, detailed and useful theories relate these entities to one another in terms of fundamental laws. Together these features and laws explain a great variety of complex and subtle phenomena.

It is widely believed that physics provides a complete catalogue of the universe's fundamental features and laws. As physicist Steven Weinberg puts it in his 1992 book *Dreams of a Final Theory*, the goal of physics is a "theory of everything" from which all there is to know about the universe can be derived. But Weinberg concedes that there is a problem with consciousness. Despite the power of physical theory, the existence of consciousness does not seem to be derivable from physical laws. He defends physics by arguing that it might eventually explain what he calls the objective correlates of consciousness (that is, the neural correlates), but of course to do this is not to explain consciousness itself. If the existence of consciousness cannot be derived from physical laws, a theory of physics is not a true theory of everything. So a final theory must contain an additional fundamental component.

Toward this end, I propose that conscious experience be considered a fundamental feature, irreducible to anything more basic. The idea may seem strange at first, but consistency seems to demand it. In the 19th century it turned out that electromagnetic phenomena could not be explained in terms of previously known principles. As a consequence, scientists introduced electromagnetic charge as a new fundamental entity and studied the associated fundamental laws. Similar reasoning should apply to consciousness. If existing fundamental theories cannot encompass it, then something new is required.

Where there is a fundamental property, there are fundamental laws. In this case, the laws must relate experience to elements of physical theory. These laws will almost certainly not interfere with those of the physical world; it seems that the latter form a closed system in their own right. Rather the laws will serve as a bridge, specifying how experience depends on underlying physical processes. It is this bridge that will cross the explanatory gap.

Thus, a complete theory will have two components: physical laws, telling us about the behavior of physical systems from the infinitesimal to the cosmological, and what we might call psychophysical laws, telling us how some of those systems are associated with conscious experience. These two components will constitute a true theory of everything.

Searching for a Theory

Supposing for the moment that they exist, how might we uncover such psychophysical laws? The greatest hindrance in this pursuit will be a lack of data. As I have described it, consciousness is subjective, so there is no direct way to monitor it in others. But this difficulty is an obstacle, not a dead end. For a start, each one of us has access to our own experiences, a rich trove that can be used to formulate theories. We can also plausibly rely on indirect information, such as subjects' descriptions of their experiences. Philosophical arguments and thought experiments also have a role to play. Such methods have limitations, but they give us more than enough to get started.

These theories will not be conclusively testable, so they will inevitably be more speculative than those of more conventional scientific disciplines. Nevertheless, there is no reason they should not be strongly constrained to account accurately for our own first-person experiences, as well as the evidence from subjects' reports. If we find a theory that fits the data better than any other theory of equal simplicity, we will have good reason to accept it. Right now we do not have even a single theory that fits the data, so worries about testability are premature.

We might start by looking for high-level bridging laws, connecting physical processes to experience at an everyday level. The basic contour of such a law might be gleaned from the observation that when we are conscious of something, we are generally able to act on it and speak about it - which are objective, physical functions. Conversely, when some information is directly available for action and speech, it is generally conscious. Thus, consciousness correlates well with what we might call "awareness": the process by which information in the brain is made globally available to motor processes such as speech and bodily action.

The notion may seem trivial. But as defined here, awareness is objective and physical, whereas consciousness is not. Some refinements to the definition of awareness are needed, in order to extend the concept to animals and infants, which cannot speak. But at least in familiar cases, it is possible to see the rough outlines of a psychophysical law: where there is awareness, there is consciousness, and vice versa.

To take this line of reasoning a step further, consider the structure present in the conscious experience. The experience of a field of vision, for example, is a constantly changing mosaic of colors, shapes and patterns and as such has a detailed geometric structure. The fact that we can describe this structure, reach out in the direction of many of its components and perform other actions that depend on it suggests that the structure corresponds directly to that of the information made available in the brain through the neural processes of awareness.



COLOR WHEEL arranges hues so that ones experienced as similar are closest. Nearby colors also correspond to perceptual representations in the brain.

Similarly, our experiences of color have an intrinsic three-dimensional structure that is mirrored in the structure of information processes in the brain's visual cortex. This structure is illustrated in the color wheels and charts used by artists. Colors are arranged in a systematic pattern - red to green on one axis, blue to yellow on another, and black to white on a third. Colors that are close to one another on a color wheel are experienced as similar. It is extremely likely that they also correspond to similar perceptual representations in the brain, as part of a system of complex three-dimensional coding among neurons that is not yet fully understood. We can recast the underlying concept as a principle of structural coherence: the structure of conscious experience is mirrored by the structure of information in awareness, and vice versa.

Another candidate for a psychophysical law is a principle of organizational invariance. It holds that physical systems with the same abstract organization will give rise to the same kind of conscious experience, no matter what they are made of. For example, if the precise interactions between our neurons could be duplicated with silicon chips, the same conscious experience would arise. The idea is somewhat controversial, but I believe it is strongly supported by thought experiments describing the gradual replacement of neurons by silicon chips. The remarkable implication is that consciousness might someday be achieved in machines.

Information: Physical and Experiential

The ultimate goal of a theory of consciousness is a simple and elegant set of fundamental laws, analogous to the fundamental laws of physics. The principles described above are unlikely to be fundamental, however. Rather they seem to be high-level psychophysical laws, analogous to macroscopic principles in physics such as those of thermodynamics or kinematics. What might the

underlying fundamental laws be? No one knows, but I don't mind speculating.

I suggest that the primary psychophysical laws may centrally involve the concept of information. The abstract notion of information, as put forward in the 1940s by Claude E. Shannon of the Massachusetts Institute of Technology, is that of a set of separate states with a basic structure of similarities and differences between them. We can think of a 10-bit binary code as an information state, for example. Such information states can be embodied in the physical world. This happens whenever they correspond to physical states (voltages, say), the differences between which can be transmitted along some pathway, such as a telephone line.

We can also find information embodied in conscious experience. The pattern of color patches in a visual field, for example, can be seen as analogous to that of the pixels covering a display screen. Intriguingly, it turns out that we find the same information states embedded in conscious experience and in underlying physical processes in the brain. The three-dimensional encoding of color spaces, for example, suggests that the information state in a color experience corresponds directly to an information state in the brain. We might even regard the two states as distinct aspects of a single information state, which is simultaneously embodied in both physical processing and conscious experience.

A natural hypothesis ensues. Perhaps information, or at least some information, has two basic aspects: a physical one and an experiential one. This hypothesis has the status of a fundamental principle that might underlie the relation between physical processes and experience. Wherever we find conscious experience, it exists as one aspect of an information state, the other aspect of which is embedded in a physical process in the brain. This proposal needs to be fleshed out to make a satisfying theory. But it fits nicely with the principles mentioned earlier - systems with the same organization will embody the same information, for example - and it could explain numerous features of our conscious experience.

The idea is at least compatible with several others, such as physicist John A. Wheeler's suggestion that information is fundamental to the physics of the universe. The laws of physics might ultimately be cast in informational terms, in which case we would have a satisfying congruence between the constructs in both physical and psychophysical laws. It may even be that a theory of physics and a theory of consciousness could eventually be consolidated into a single grander theory of information.

A potential problem is posed by the ubiquity of information. Even a thermostat embodies some information, for example, but is it conscious? There are at least two possible responses. First, we could constrain the fundamental laws so that only some information has an experiential aspect, perhaps depending on how it is physically processed. Second, we might bite the bullet and allow that all information has an experiential aspect - where there is complex information processing, there is complex experience, and where there is simple information processing, there is simple experience. If this is so, then even a thermostat might have experiences, although they would be much simpler than even a basic color experience, and there would certainly be no accompanying emotions or thoughts. This seems odd at first, but if experience is truly fundamental, we might expect it to be widespread. In any case, the choice between these alternatives should depend on which can be integrated into the most powerful theory.

Of course, such ideas may be all wrong. On the other hand, they might evolve into a more powerful proposal that predicts the precise structure of our conscious experience from physical processes in our brains. If this project succeeds, we will have good reason to accept the theory. If it fails, other avenues will be pursued, and alternative fundamental theories may be developed. In this way, we may one day resolve the greatest mystery of the mind.

Dancing Qualia in a Synthetic Brain [Box]

Whether consciousness could arise in a complex, synthetic system is a question many people find intrinsically fascinating. Although it may be decades or even centuries before such a system is built, a simple thought experiment offers strong evidence that an artificial brain, if organized appropriately, would indeed have precisely the same kind of conscious experiences as a human being.

Consider a silicon-based system in which the chips are organized and function in the same way as the neurons in your brain. That is, each chip in the silicon system does exactly what its natural analogue does and is interconnected to surrounding elements in precisely the same way. Thus, the behavior exhibited by the artificial system will be exactly the same as yours. The crucial question is: Will it be conscious in the same way that you are?

Let us assume, for the purpose of argument, that it would not be. (Here we use a reasoning technique known as *reductio ad absurdum*, in which the opposite hypothesis is assumed and then shown to lead to an untenable conclusion.) That is, it either has different experiences - an experience of blue, say, when you are seeing red - or no experience at all. We will consider the first case; the reasoning proceeds similarly in both cases.

Because chips and neurons have the same function, they are interchangeable, with the proper interfacing. Chips therefore can replace neurons, producing a continuum of cases in which a successively larger proportion of neurons are replaced by chips. Along this continuum, the conscious experience of the system will also change. For example, we might replace all the neurons in your visual cortex with an identically organized version made of silicon. The resulting brain, with an artificial visual cortex, will have a different conscious experience from the original: where you had previously seen red, you may now experience purple (or perhaps a faded pink, in the case where the wholly silicon system has no experience at all).

Both visual cortices are then attached to your brain, through a two-position switch. With the switch in one mode, you use the natural visual cortex; in the other, the artificial cortex is activated. When the switch is flipped, your experience changes from red to purple, or vice versa. When the switch is flipped repeatedly, your experiences 'dance' between the two different conscious states (red and purple), known as qualia.

Because your brain's organization has not changed, however, there can be no behavioral change when the switch is thrown. Therefore, when asked about what you are seeing, you will say that nothing has changed. You will hold that you are seeing red and have seen nothing but red, even though the two colors are dancing before your eyes. This conclusion is so unreasonable that it is best taken as a *reductio ad absurdum* of the original assumption-that an artificial system with identical organization and functioning has a different conscious experience from that of a neural brain. Retraction of the assumption establishes the opposite: that systems with the same organization have the same conscious experience. -D.J.C.

Why Neuroscience May Be Able to Explain Consciousness [Box]

by Francis Crick and Christof Koch

We believe that at the moment the best approach to the problem of explaining consciousness is to concentrate on finding what is known as the neural correlates of consciousness: the processes in the brain that are most directly responsible for consciousness. By locating the neurons in the cerebral cortex that correlate best with consciousness, and figuring out how they link to neurons elsewhere in the brain, we may come across key insights into what David J. Chalmers calls the hard problem: a full accounting of the manner in which subjective experience arises from these cerebral processes.

We commend Chalmers for boldly recognizing and focusing on the hard problem at this early stage, although we are not as enthusiastic about some of his thought experiments. As we see it, the hard problem can be broken down into several questions: Why do we experience anything at all? What leads to a particular conscious experience (such as the blueness of blue)? Why are some aspects of subjective experience impossible to convey to other people (in other words, why are they private)? We believe we have an answer to the last problem and a suggestion about the first two, revolving around a phenomenon known as explicit neuronal representation.

What does 'explicit' mean in this context? Perhaps the best way to define it is with an example. In response to the image of a face, say, ganglion cells fire all over the retina, much like the pixels on a television screen, to generate an implicit representation of the face. At the same time, they can also respond to a great many other features in the image, such as shadows, lines, uneven lighting and so on. In contrast, some neurons high in the hierarchy of the visual cortex respond mainly to the face or even to the face viewed at a particular angle. Such neurons help the brain represent the face in an explicit manner. Their loss resulting from a stroke or some other injury, leads to prosopagnosia, an individual's inability to recognize familiar faces, consciously-even his or her own, although the person can still identify a face as a face. Similarly, damage to other parts of the visual cortex can cause someone to lose the ability to experience color, while still seeing in shades of black and white, even though there is no defect in the color receptors in the eye.



KANIZSA TRIANGLE stimulates neurons that code explicitly for such illusory contours.

At each stage, visual information is re-encoded, typically in a semi-hierarchical manner. Retinal ganglion cells respond to areas of light. Neurons in the primary visual cortex are most adept at responding to lines to edges; neurons higher up might prefer a moving contour. Still higher are those that respond to faces and other familiar objects. On top are those that project to pre-motor and motor structures in the brain, where they fire the neurons that initiate such actions as speaking or avoiding an oncoming automobile.

Chalmers believes, as we do, that the subjective aspect of an experience must relate closely to the firing of the neurons corresponding to those aspects (the neural correlates). He describes a well-known thought experiment, constructed around a hypothetical neuroscientist, Mary, who specializes in color perception but has never seen - a color. We believe the reason Mary does not know what it is like to see a color, however, is that she has never had an explicit neural representation of a color in her brain, only of the words and ideas associated with colors.

In order to describe a subjective visual experience, the information has to be transmitted to the motor output stage of the brain, where it becomes available for verbalization or other actions. This transmission always involves re-encoding the information, so that the explicit information expressed by the motor neurons is related, but not identical, to the explicit information expressed by the neurons associated with color experience, at some level in the visual hierarchy.

It is not possible, then, to convey with words and ideas the exact nature of a subjective experience. It is possible, however, to convey a difference between subjective experiences-to distinguish between red and orange, for example. This is possible because a difference in high-level visual cortical area will still be associated with a difference in the motor stages. The implication is that we can never explain to other people the nature of any conscious experience, only its relation to other ones.

The other two questions, concerning why we have conscious experiences and what leads to specific ones, appear more difficult. Chalmers proposes that they require the introduction of 'experience' as a fundamental new feature of the world, relating to the ability of an organism to process information. But which types of neuronal information produce consciousness? And what makes a certain type of

information correspond to the blueness of blue, rather than the greenness of green? Such problems seem as difficult as any in the study of consciousness.

We prefer an alternative approach, involving the concept of 'meaning.' In what sense can neurons that explicitly code for a face be said to convey the meaning of a face to the rest of the brain? Such a property must relate to the cell's projective field a pattern of synaptic connections to neurons that code explicitly for related concepts. Ultimately, these connections extend to the motor output. For example, neurons responding to a certain face might be connected to ones expressing the name of the person whose face it is and to others for her voice, memories involving her and so on. Such associations among neurons must be behaviorally useful, in other words, consistent with feedback from the body and the external world.

Meaning derives from the linkages among these representations with others spread throughout the cortical system in a vast associational network, similar to a dictionary or a relational database. The more diverse these connections, the richer the meaning. If, as in our previous example of prosopagnosia, the synaptic output of such face neurons were blocked, the cells would still respond to the person's face, but there would be no associated meaning and, therefore, much less experience. A face would be seen but not recognized as such.

Of course, groups of neurons can take on new functions, allowing brains to learn new categories (including faces) and associate new categories with existing ones. Certain primitive associations, such as pain, are to some extent inborn but subsequently refined in life.

Information may indeed be the key concept, as Chalmers suspects. Greater certainty will require consideration of highly parallel streams of information, linked-as are neurons-in complex networks. It would be useful to try to determine what features a neural network (or some other such computational embodiment) must have to generate meaning. It is possible that such exercises will suggest the neural basis of meaning. The hard problem of consciousness may then appear in an entirely new light. It might even disappear.

The Author

David J. Chalmers studied mathematics at Adelaide University and as a Rhodes Scholar at the University of Oxford, but a fascination with consciousness led him into philosophy and cognitive science. He has a Ph.D. in these fields from Indiana University and is currently in the department of philosophy at the University of California Santa Cruz. Chalmers has published numerous articles on artificial intelligence and the philosophy of mind. His book *The Conscious Mind*, which elaborates many of the ideas in this article, is forthcoming from Oxford University Press.

Further Reading

Absent Qualia, Fading Qualia, Dancing Qualia. David J. Chalmers in *Conscious Experience*. Edited by Thomas Metzinger. Ferdinand Schoningh, 1995.

Explaining Consciousness: The Hard Problem. Special issue of Journal of Consciousness Studies, Vol. 2, No. 3; Autumn 1995.

The Nature of Consciousness: Philosophical and Scientific Debates. Edited by Ned Block, Owen Flanagan and Güven Güzeldere. MIT Press (in press).

Consciousness and Cognition[*]

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

*[[I wrote this paper in January of 1990, but did not publish it because I was never entirely happy with it. My ideas on consciousness were in a state of flux, ultimately evolving into those represented in my book *The Conscious Mind* (Oxford University Press, 1996). I now think that some parts of this paper are unsatisfactory, especially the positive theory outlined at the end, although a successor to that theory is laid out in my recent work. Nevertheless, I think the paper raises issues that need to be addressed.]]

Introduction

The problem of consciousness is perhaps the largest outstanding obstacle in our quest to scientifically understand reality. The science of physics is not yet complete, but it is well-understood. The science of biology has explained away most of the mysteries surrounding the nature of life. Where there are gaps in our understanding of these fields, the gaps do not seem intractable; we at least have some idea of the direction in which solutions might lie. In the science of mind, things are not quite so rosy. Much progress is being made in the study of *cognition*, but *consciousness* itself is as much of a problem as it ever was.

The term "consciousness" often serves as a catch-all for all that is mysterious about mentality. When using the term, one must therefore be careful not to collapse important distinctions. The most important distinction in the study of consciousness bears upon the approach we take to studying it: we may take either the *first-person* or the *third-person* approach. We might say that the third-person approach treats consciousness as a scientific problem, while the first-person approach treats it as a metaphysical problem. These two different viewpoints that we can adopt lead to very different treatments of the phenomenon, each of which can supply crucial insights. But the reconciliation of the viewpoints seems to be a difficult matter. Frequently, proponents of the two approaches seem to be talking past each other. In this paper, I will try to explicate carefully the relation between the two approaches, and to argue that they are not so irreconcilable as they might seem.

The third-person approach

The third-person approach has much to recommend it. On this view, consciousness is treated as a problem of science, just like heat, life, or nuclear physics, and amenable to the same methods of inquiry. Researchers in the new discipline of cognitive science have had much success in taking this approach to

the study of mental processes in general. It is natural to hope that it might shed some light on the problems of consciousness, too.

The raw materials of cognitive science are much the same as those of any science - data gathered from external observation. These data can take a number of different forms. The most obvious kind of observable data are behavioral, and the study of human behavior is concentrated in the field of psychology. Another source of data, less accessible than behavior but extremely useful for the third-person approach, is observation of brain structure and function. The field of neuroscience is beginning to have significant success in explaining how the brain supports cognition. A third, less direct method in cognitive science is that of cognitive modeling - the construction of models, usually computational, that cohere to some extent with behavioral or neurophysiological data. The field of artificial intelligence is concerned with producing such models, and researchers in psychology and neuroscience often use computational models in their work.

This tripartite investigation of behavior, brain function and cognitive models has led to a significant increase in our understanding of diverse aspects of cognition, such as vision, memory and language comprehension. It is not surprising that these third-person methods might also be used to investigate consciousness. Such an approach to consciousness is urged by Dennett (1978; 1991), who goes so far as to sketch cognitive models for the phenomenon. Another third-person approach has been taken by Jackendoff (1987) (although Jackendoff explicitly recognizes that a third-person approach will not necessarily tell the whole story). Besides these philosophers, some psychologists (e.g., Johnson-Laird 1983) and computer scientists (e.g., Hofstadter 1979) have tried their hand at reining in the mysteries of consciousness with third-person accounts.

At the heart of the third-person approach is the philosophical position of *functionalism*. Roughly stated, this is the view that the correct way to understand mental processes is to perform causal analyses, revealing the abstract causal structure behind brain function. Such causal structure can be understood in objective terms, and duplicated in quite different materials (including, perhaps, in computers). Central to this approach is the view of *mind as system*. Insofar as this view addresses consciousness, it takes it to be a particular aspect of the function of a complex system; for example, it might be taken to be that process whereby a system is able to scan its own processing. The view of mind as system is certainly attractive, and it is a view that I share. The question is whether this approach taken alone is enough to explain the deepest mysteries of consciousness.

The first-person approach

Despite the attractiveness of the third-person approach, there is a feeling that it is side-stepping the really hard problems. The truly difficult questions only seem to arise when we take the first-person approach - when we consider, as Nagel famously put it, what it is like to *be* who we are, and what it might be like to be something quite different: another human, a bat, or a computational system. It is with these subjective questions that the deepest questions arise. Why is being me like anything at all? And why is it the way that it is? These questions are the real content of the problem of consciousness. There may be other senses of the term "consciousness", but in the sense in which it is most commonly used, to refer to the

real mysteries of mentality, it is these first-person questions that are being raised.

It is still easy to fall into confusion or to equivocate when talking of "consciousness", so here I will divide the first-person problem into three parts: the problems of sensory qualia, subjective mental content, and the existence of subjective experience.

(1) The problem of sensory qualia

Qualia are the qualitative aspects of our mental states, most obviously of our sensations. The paradigm qualia are those of color sensations; other favorites are the taste of chocolate, the sound of middle C, pleasure and pain. All of these are poorly understood. When we look at a red patch, this sets off a particular pattern of neural firings in our brain. Why should this physical process be accompanied by a rich, subjective sensation? Given that it is accompanied by a sensation, why is it *this* sort of sensation (the red sort) rather than *that* sort (the green sort)? There are two issues here: why qualia exist at all, and why particular qualia accompany particular processes. Is the correspondence of qualia to processes arbitrary, or is there some systematicity that we do not understand?

Jackson (1982) has provided the most recent reminder of the qualia mystery, with a sharpening of the argument of Nagel (1974) before him. A future scientist, living in a time when neuroscience is completely understood, might learn everything there is to know about physical brain-processes. But if she has lived all her life in a black-and-white room, she will still not know what it is like to see red; when she sees red for the first time, she will learn something. It seems that the third-person approach, at least as currently understood, cannot tell us about the nature of qualia.

(2) The problem of subjective mental content.

When I think about a lion, something takes place in my subjective experience that has something to do with lions. Again, a straight physical account gives no reason to believe that such an experience should take place. What should a pattern of neural firings have to do with *lions*? But somehow, my thoughts are about something; they have subjective mental content. It is easy to make *attributions* of mental content to a system, justified perhaps by causal relations with the external world, but for subjective mental content we need something stronger. We need brain-states to carry *intrinsic* content, independent of our systems of external attribution; there must be a natural (in the strongest sense, i.e., defined by nature) mapping from physical state to content.

The problem of subjective mental content is not entirely different in kind from that of sensory qualia - the experience of content is itself qualitative, in a way. The main difference is that sensory qualia usually arise during external perception, whereas this sort of mental content arises during thought. (There is also a third-person problem of mental content, which has been raging for years, centering on the question of how we can assign propositional attitudes, such as beliefs and desires concerning the world, to systems and persons. In some ways this is an easier problem, as it may rely on human-defined systems of attributions; these contents may have the status of theoretical entities, rather than states that are presented

to us directly. In other ways, the first-person problem is easier, as it may not have to deal with the problem of *reference*. When I think of a lion, my phenomenology bears *some* relation to a lion, but the relationship seems more like shared pattern than reference.)

(3) *The existence of subjective experience*

The two items above are concerned with the *nature* of our subjective states - why they are one way rather than another. But it is just as deep a problem why subjective states should exist in the first place. Why should it be like *anything* to be me? If I did not *know* that subjective states existed, it would seem unreasonable to postulate them. This is perhaps the deepest question of all, and no current theory has come close to dealing with it.

Not many people believe in zombies - humans with normal behavior but without any subjective mental states. These may be logically possible, but it seems implausible that there could be such things in the actual world. At least some people believe there could be functional zombies, however: beings which duplicate the functional organization of humans, perhaps computationally, without being conscious at all (e.g. Searle 1980, Block 1980). The question "what sort of entities can be subjects of experience?" is of great popular interest. For example, is every entity exhibiting intelligent behavior conscious? Could an appropriately programmed computer be conscious? I will argue, by combining first-person and third-person considerations, that the possible existence of functional zombies is implausible.

There are some other commonly-raised first-person problems not explicitly listed above. The problem of *self-consciousness* (or self-awareness) I take to be a subset of the problem of awareness, of which the difficult aspects are covered by (1) and (2). The problem of *personal identity* is a separate issue and a very deep one; but Parfit's exhaustive analysis (1986), which combines the first-person and third-person approaches to great effect, gives reason to believe that our first-person intuitions here may be mistaken. It is the three problems listed above that seem to be the residual content of the traditional mind-body problem. I will be using the term "consciousness" broadly to cover the phenomena of all of these problems. If you prefer, replace every occurrence of "consciousness" with "the subjective experience of qualia and mental content."

One difficulty with talking about first-person problems is that for every first-person (conscious) mental state, there is a corresponding third-person (functionally definable) mental state. (Perhaps there are not two different mental states, but simply two different ways of viewing one state; this is unclear. In any event, it is uncontroversial that for every subjective mental event there is a corresponding physical event; the physical event may be viewed via functional abstraction as a third-person mental event). For every subjective *sensation* there corresponds an objectively characterizable *perception*. This dichotomy in (ways of looking at) mental states makes things a little confusing, but it will be useful later on.

The relationship between the approaches

Of course, the first-person approach and the third-person approach are not entirely independent of each

other. Indeed, I will be arguing that much insight can come through considering them not singly, but juxtaposed. Before doing this, it is necessary to enumerate a few facts about the two approaches.

(1) The third-person approach is sufficient, in principle, to yield a complete explanation of human behavior.

This follows from the explanatory completeness of physical laws - a hypothesis that is relatively uncontroversial. Physical phenomena have physical explanations, and behavior is a physical phenomenon. There is no reason to believe that human behavior is any different in this regard from hurricanes, rivers or stars. With a complete understanding of the laws of physics, and a good understanding of the way physical entities combine into systems, behavior will be understood.

This is an extremely strong conclusion, of course, and it goes a long way towards explaining the dominance of the third-person approach to the study of mentality. Cognitive science is doing its best to break down the removable barrier implied by the "in principle", and there is no reason to believe that within a few centuries it should not succeed. If it does not, it will probably be because of problems of complexity rather than any deep metaphysical barrier. We might trivially restate (1) as

(1a) The first-person approach is not needed to explain anything about human behavior.

Of course, the first-person approach may well prove to be a useful short-cut in explaining behavior, and a powerful tool, but the point is that it is not *necessary* - it is in principle dispensable, as far as behavior is concerned.

(2) The third-person approach (as currently conceived) is not sufficient to explain all first-person phenomena.

Statements (1) and (1a) make a powerful argument for the primacy of the third-person approach. Nevertheless, this approach cannot tell the whole story. Consciousness, the subjective experience of qualia and mental content, is simply not explained by a third-person account. No amount of neuroscience and cognitive modeling explains the qualitative nature of a sensation of red, or even why such a subjective sensation should exist.

This statement is surely the most natural view, but it is not undisputed. Those who have spoken the loudest in dispute of (2) include Dennett (1982, 1988, 1991) and P. M. Churchland (1986). I believe their arguments to be flawed, but will not argue the issue here; it has been eloquently argued by others. Those who disbelieve (2) may choose to stop reading here, then, in the belief that consciousness is not really a special problem. They might however read on, to find an account of first-person problems which is motivated by third-person issues at every point, or at worst to find a *reductio ad absurdum* of the first-person position. The remainder - those who believe that the mind-body problem is still a real problem - may continue undisturbed.

It should be noted that (2) refers to the third-person approach *as currently conceived* - that is, the study of the brain as a physical system with the usual physical ontology, abetted by cognitive models and functional analyses. I would not like to rule out the possibility of an eventual "objective phenomenology", to use Nagel's term, but it is difficult to see what it would look like. At the least, new constructs would be needed, and our ontology would need to be expanded. To get from the physical and the functional to the subjective, we would need *metaphysical bridging principles*. It does not seem impossible, however, that these could be stated from a third-person viewpoint. The beginnings of such an approach will be outlined later in this paper. In the meantime, we can restate (2) as

(2a) The third-person approach is not sufficient, in the absence of extra metaphysical bridging principles, to explain all first-person phenomena.

The need for such a bridge is put nicely by Jackendoff (1987), who says that computational/functional accounts of mind may have solved the Mind-Body Problem, but what we need to do now is solve the "Mind-Mind Problem". This would provide a bridge from the "computational mind" to the "phenomenological mind" (in the terms I have used, from third-person mental events to first-person mental events). That computational and functional accounts may have already got us halfway there is a point that will resurface later.

(3) Our claims about consciousness are a fact of human behavior.

This is trivially true. (If you interpret "claim" as an irreducibly subjective term, then replace it by "utterance" or something more neutral.) When I say "I'm conscious, and I'm totally baffled by it," that is a fact of human behavior. When somebody writes "The ineffable sensation of red is inexplicable by physical premises", that is a fact of human behavior. This is harmless enough, until combined with (1) to yield

(4) The third-person approach is in principle sufficient to explain our claims about consciousness

or worse,

(4a) The first-person approach is not needed to explain our claims about consciousness.

Everything that we say about consciousness is, in principle, amenable to the usual kind of physical/functional analysis. There is no need to appeal to *any* mysterious metaphysical constructs to explain the things we say. On this account, the last 2000 years of debate over the Mind-Body Problem would have gone exactly the same in the absence of subjective experience (if such a thing were possible). In particular, our sense of bafflement about the first-person problems manifests itself in numerous third-person ways - the things we say, the things we write, even the things we *think*, if regarded in terms of neural processes - and is thus amenable to the methods of cognitive science[*]. The correct cognitive-science explanation of why we claim to be conscious might be a little while coming, but there should be

no deep metaphysical problems. This leads us to an important principle:

*[[McGinn (1989) requests an explanation of why it is so difficult to clearly articulate the the problems of consciousness. I believe that the reason is closely tied up with (4) and (4a). Everything we can say about consciousness is physically caused, and somehow, at the back of our minds, we are aware of this. So when the words come out of our mouths, we are aware of how easy it might be to explain our claims without invoking any great mysteries. Further, our words can easily be taken as referring to third-person mental states: perception rather than sensation, and so on. Our claims seem very inadequate to do justice to the real phenomenon.]]

The Mystery Principle: Consciousness is mysterious. Claims about consciousness are not.

This can equivalently be stated as

The Surprise Principle: Although consciousness should surprise us, claims about consciousness should not.

Most people will agree that consciousness is a surprising phenomenon. If it were not for the fact that first-person experience was a brute fact presented to us, there would seem to be no reason to predict its existence. All it does is make things more complicated. By contrast, the things we say about consciousness are common-or-garden cognitive phenomena. Somebody who knew enough about brain structure would be able to immediately predict the likelihood of utterances such as "I *feel* conscious, in a way that it seems no physical object could be", or even Descartes' "Cogito ergo sum". In short, our bafflement about consciousness can be understood purely as a problem for cognitive science. There are very rich pickings awaiting anybody who takes this path. I believe that it is not as difficult a path as it might seem, and towards the end of this paper I will give a very brief account of why we might expect such bafflement.[*]

*[[There are other reasons why claims about consciousness should not surprise us. It seems plausible that any system which (in a purely functional sense) "perceives" aspects of the world and further has sophisticated linguistic capacity might profess perplexity about the nature of its perception; one might expect puzzlement at both the "qualitative" and "indexical" aspects. And if such a system has the capacity to "think about what it is thinking about" (a functional notion, once again), it might very well claim full-blown self-consciousness. No metaphysics needs to be invoked in our analysis of such a situation. (Computer A: "I know I'm just a collection of digital circuits, but I can't explain this strange *feeling* I have of having thoughts of my own." Computer B: "Yes, I'm inclined towards dualism myself.") This point deserves further development.]]

The situation with which we are now faced seems to border on the absurd. If everything we say, write and even *think* about consciousness is explainable by cognitive science, why is there a need to posit any great mystery about consciousness in the first place? Why not concede that the appearance of great metaphysical mystery is simply an illusion, and regard consciousness itself as a cognitive process, on much the same level as memory or learning? It becomes a third-person high-level term, playing a similar role in cognitive science as "heat" does in thermodynamics. As for this weird "first-person experience", then insofar as it is not a purely cognitive phenomenon, it simply does not exist. This is the approach taken by Dennett and others, and there are powerful arguments in its favor, as evidenced above. But this

paper is based on the premise that the mind-body problem is still a real problem. We are not yet prepared to concede that (2) is false, and that the third-person approach as currently conceived is sufficient to explain all of the first-person mysteries. As first-person loyalists, we have to bite this bullet and struggle on, trying to produce a coherent account of the matter. I believe that the pressures on a first-person account produced by (4) and (4a), and the third-person constraints that are thus acquired, are of great value in shaping a solution that might be acceptable to partisans of both the first-person and third-person approaches.

This is the low point for the first-person partisan. This is as bad as it gets. But we have to confront the problem posed by (4) and (4a) head on. This has rarely been done in the literature. "Qualia freaks" (Jackson's term) have generally been content to argue for the autonomy of first-person phenomena, from independent considerations, or more frequently to simply assume it. But if we do not take (4) and (4a) seriously, then any account that we produce is open to question. The question which faces us, put slightly differently, is

(4b) How can first-person experience have any bearing on what we say about first-person experience?

or more simply

(4c) How can consciousness have any bearing on what we say about consciousness?

I believe that in the proper understanding of this question lies the seed of a solution to the mind-body problem. All the same, it is surprising how rarely the question has been considered in the literature. (The more general question of how the mental can affect the physical of course a perennial favorite. Question (4c) is much more specific, and raises a different set of problems.) It has been touched on occasionally in various forms, however, with various different reactions. The choices in answering the question spread themselves out as follows.

(A) We might reject premise (1), and deny that the third-person approach can even give a full account of human *behavior*. This is the approach taken by Elitzur (1989), who argues from the fact that we talk about consciousness to the conclusion that consciousness plays an active role in determining human behavior, and deduces that the laws of physics must therefore play an incomplete role in determining behavior. This is certainly an interesting argument, but it is implausible and should be regarded as a last resort. The laws of physics are too important to us to be given up so easily. (A similar account of the matter would presumably be given by interactionist dualists from Descartes to Eccles.)

(B) We might answer (4c) by saying, simply, "It doesn't." This is the answer that might be given by Jackson (1982), who argues that qualia are completely epiphenomenal, playing no causal role (although he does not explicitly consider utterances about consciousness). Now, epiphenomenal qualia are already unparsimonious under general considerations, but considered in the context of question (4c) their implausibility increases. Are we to say that our claims about consciousness are completely independent

of the fact of our consciousness? That any similarity manifested between our claims and the actual facts are mere coincidence? While this conclusion cannot be completely ruled out, it certainly seems inelegant. It fails what I will call, later, the "Coherence Test" - its explanations of (1) consciousness and (2) claims about consciousness fail completely to cohere with each other.

I think there may nevertheless be a grain of truth in epiphenomenalism, but it needs to be spelt out more carefully. If we can explicate just *how* consciousness is an epiphenomenon of physical processing, then question (4c) may admit a more parsimonious answer. I will be defending a kind of "double aspect" theory that bears some similarity to epiphenomenalism, while at the same time being more palatable, I hope, to those with a naturalistic view of the world.

(C) We may deny premise (2), and argue that in fact the third-person approach tells us everything there is to know. By denying that consciousness is essentially a first-person phenomenon, we deny that question (4c) poses any problem. If there is anything at all called "consciousness" that plays a causal role, then it does so in exactly the same way that centers of gravity play a causal role in physics, or that temperature plays a causal role in thermodynamics: as a convenient third-person abstraction. This is presumably the approach that the likes of Dennett and Churchland would take.

The consideration of questions like (4c) has been used as an explicit argument for the third-person approach in a few places. For example, Foss (1989), arguing against the Nagel/Jackson arguments, argues that a "Super Neuroscientist" could in principle know everything that a being would say about its qualia-experience, and everything that it *might* say. From this, he draws the conclusion that the Super Neuroscientist would know everything that there is to know about qualia.

Clearly questions like (4c) are a goldmine for third-person arguments, but these analyses seem too simplistic. They "solve" the Mind-Body Problem by denying that there is a problem. This runs counter to our premise that there is a substantial problem to be solved. Nevertheless, I believe there is a grain of truth in these arguments. There is perhaps some sense in which qualia are physical. There is perhaps some sense in which consciousness is understandable from the third-person viewpoint. What must be explicated is the nature of this sense. The arguments above do not solve the Mind-Body Problem, but they do indicate that it might be soluble. However, there is more work to be done. A flat statement of physicalism is in no sense a complete solution. We need to know *how* physicalism could be true. The nature of the work that remains to be done will be outlined shortly.

Before we leave this section, we should note that a well-known argument of Shoemaker (1975) bears some resemblance to the arguments in (C). Shoemaker argues against the possibility of "Absent Qualia", or functional zombies. Some have argued that this possibility refutes functionalism, by demonstrating that we might duplicate any given functional process without any accompanying subjective experience. Shoemaker responds by arguing that we *know* about qualia; our knowledge is caused by the existence of these qualia; therefore qualia play some causal role. The conclusion is that Absent Qualia are impossible, as removing qualia would change the causal structure. This is generally construed as an argument for the third-person approach (for functionalism, in particular). Nevertheless, I am sympathetic with this argument and with functionalism, and believe that Absent Qualia are impossible. However, what needs to

be explained is *why* Absent Qualia are impossible, and *how* qualia could play a causal role. This question is not addressed by functionalism. Again, more work remains to be done. Both functionalism and epiphenomenalism have plausible aspects, but neither gives a complete account of the problems of consciousness. I will suggest that a correct theory of consciousness shares aspects of both functionalism and epiphenomenalism, and that the correct answer to question (4c) lies somewhere between (B) and (C).

The Coherence Test

The Coherence Test is a test that any completed theory of mind must pass. It is motivated directly by questions like (4c).

The Coherence Test:

A completed theory of mind must provide

(C1) An account of *why* we are conscious.

(C2) An account of *why* we *claim* to be conscious; *why* we *think* we are conscious.

Further (C3), accounts (C1) and (C2) must *cohere* with each other.

It should be noted that (C1) is a completely first-person matter; it involves answering the metaphysical question of how subjective experience is possible. (C2), on the other hand, is understandable on strictly third-person terms - it is, in principle, a matter for cognitive science. (If, again, in your vocabulary "think" and "claim" are irreducibly subjective terms, then replace them by appropriate third-person terms that deal with behavior and brain function.) Another way in which to phrase (C2) might be "An account of why consciousness seems so strange to us." Viewed appropriately, this is also in principle understandable by a cognitive science approach. Anyway, it is the coherence condition (C3) that is the most important here. This ensures that completely independent accounts under (C1) and (C2) will be ruled out; that the accounts must bear a substantial relationship to each other. This provides a vital link between the first-person and third-person approaches.

Another way of putting the Coherence Test is that we need accounts of

(C1a) The mind that we experience (a first-person question); and

(C2a) The mind that the system perceives (a third-person question).

One would hope that these "minds", on any completed account, would bear a very close relationship to each other. If the relationship is not strict identity, it should at least be a very close correspondence.

It is not possible to "prove" that a correct theory of consciousness must pass the Coherence Test, but it certainly seems extremely plausible. While it is logically possible that an account of why we are conscious could be completely independent of a cognitive account of why we *believe* we are conscious, it seems inelegant and unlikely. If the Coherence Test in fact fails for a correct completed theory of mind, then we are living in a world where the things we say about consciousness and the Mind-Body Problem bear at best a coincidental relation to the way these things actually are. Faced with the possibility of this complete unreliability of our verbal reports, it would probably be better to join the third-person camp and

regard the mysterious nature of "subjective experience" as a mere illusion.

Given that we accept the Coherence Test, a correct third-person account of cognitive processes puts severe constraints on possible first-person accounts, and possibly vice versa. Of course, we have not explicitly defined the notion of coherence. Instead, I think it is better to go ahead and apply it on a case-by-case basis, and let some understanding of the notion emerge.

Applying the Coherence Test

Functionalism is a controversial doctrine when applied to first-person aspects of mentality. But for the third-person study of mental processes, it reigns supreme. Few people these days doubt that the correct way to analyze human behavior and to explain the workings of the brain is in functional terms. All remotely satisfactory accounts of learning or memory, say, explain these via causal analyses of systems with many interacting parts. Indeed, it is difficult to see what other kind of account there could be. So when it comes to explaining the things we *say* (and even *believe*) about consciousness, there is little doubt that a functional analysis will be the way to go. We may enshrine this doctrine as

(5) The fact that we claim to be conscious holds in virtue of certain functionally specifiable properties of the brain.

Of course, we are not sure exactly what those functional properties *are*, yet - this is part of the reason why functionalism alone is not a completed theory of consciousness. But we are almost certain that the properties will be functional ones, considered at the most parsimonious level. And there is no doubt that if we make the correct functional abstraction from the brain, then duplicating this functional specifications will produce the same kind of claims about consciousness.

We may now use this uncontroversial claim, together with the Coherence Test, as an argument against certain theories of mind.

Theory 1: We are conscious in virtue of low-level biochemical processes.

Such a view of mind is held by at least a few writers on the subject (e.g. Searle 1980, Block 1980). These writers have generally been arguing against functionalism as a theory of mind, or at least as a theory of the first-person aspects. Of course, almost everybody agrees that neurophysiological processes give rise to consciousness in the human case; but that view alone is compatible with functionalism, as these processes might give rise to consciousness precisely in virtue of their functional properties. Proponents of the present view take a stronger stand, suggesting that biochemical properties are *directly* responsible for consciousness. On this view, *only* a system with the relevant biochemical properties could be conscious; a silicon isomorph might lack consciousness entirely. Block, for example, claims that qualia might be a consequence of physiological facts which need not appear in a functional account. Searle thinks it very plausible that the "causal powers" of the brain (meaning those powers sufficient to cause consciousness) lie at the biochemical level.

This view, however, runs a great risk of failing the Coherence Test. If, as according to (5), our claims of consciousness are consequences of certain functional properties of the brain, then the statement that consciousness is inherently neurophysiological suggests that reason we make the claims is largely independent of the reason we are conscious. This would seem strange. A possible reply by neurophysiological partisans might be to claim that the neurophysiological locus of consciousness is specifiable functionally, albeit at a very low level. The connectionist movement in cognitive modeling has recently made popular the idea that a correct functional account may inhere at a low level of abstraction. But if we accept that neurophysiological accounts can also be functional, then this is no longer an argument against functionalism per se - a consequence that Block and Searle would be loath to accept. Insofar as our theory holds that the sources of consciousness are not specifiable functionally, then it is a theory that fails the Coherence Test. If consciousness does not arise from the functional, then our claims are independent of consciousness.

In reply to a related argument by Shoemaker (1975), Block (1980) suggests that qualia can be independent of any given functional account but still make a causal difference. But such a difference will by definition not bear on any aspect of our functional account. Given that our account is specifically constructed to include our claims of consciousness as a relevant aspect, this reply fails. If qualia are independent of this functional account, then qualia cannot make a difference to our claims. Horgan (1984) makes a similar move in defending the claim that qualia are neurophysiological. Considering the case of Martians functionally identical to us but physiologically different, he claims that the issue of whether or not they have qualia might be resolved by observing whether they manifest the same kind of puzzlement about qualia that we do. But *of course* they will manifest similar puzzlement, as this is determined by their functional architecture. Therefore such observation will tell us nothing more than we already knew.

For similar reasons, even if we hold that the underpinnings of consciousness are specifiable functionally, it may be dangerous to go to too low a level. It may well be that when an account of the processes which lead to our consciousness-claims is found, it will be specifiable at a level higher than the biochemical. If we insist that consciousness itself inheres at a lower level than this, then the source of consciousness is independent of the source of our consciousness-claims, and our theory fails the Coherence Test. (It may be possible that certain *aspects* of qualia are determined at a lower level than the level of our qualia-claims. Such aspects could not make any causal difference relevant to our claims. For instance, the difference between red-sensations and green-sensations is very difficult to articulate, and so the precise nature of these sensations might be dependent on certain low-level facts. But the *existence* of such qualia must be dependent on the higher-level functional account, if the correct theory of mind is Coherent.)

We may conclude that theories which place the source of consciousness at too low a level are implausible. A similar argument will apply to any theory which does not satisfy Functional Supervenience: the principle that replicating certain relevant functional processes, even in a different substrate, is sufficient to guarantee first-person mental states. (This is a weaker principle than the Functional State Identity Theory, which holds that mental states *are* functional states.) If consciousness exists in virtue of some fact which is quite independent of any functional account, then our claims of

consciousness do not reflect the fact of our consciousness. This argument parallels Shoemaker's argument against Absent Qualia in some ways, although unlike Shoemaker I do not wish to argue for the truth of the Functional State Identity Theory. But the argument establishes that if there could exist beings which duplicate our relevant causal structure (perhaps in a quite different substrate), but which do not possess subjective states - in other words, if there could exist functional zombies - then the correct theory of mind is fundamentally Incoherent.[*]

*[[It should be noted that the implausibility of functional zombies does not necessarily imply the impossibility of computational zombies. It just might be possible to be a functionalist without believing that computational specifications can ever capture the appropriate causal structure. (Harnad (1989), for instance, believes that the requisite function is inherently analog.) Church's Thesis diminishes the plausibility of such a position - it seems likely that a non-computational causal account always supervenes on a lower-level computational account - and in practice functionalism and computationalism tend to stand and fall together.]]

Theory 2: Epiphenomenalism.

On the face of it, this view does not come close to satisfying the Coherence Test. It holds that we are conscious in virtue of certain dualistic principles which yield mental states as a consequence of physical states. Such principles would seem to be quite independent of the relatively straightforward functional sources of our claims that we have "qualia" and "subjectivity". All parsimony is thrown out the window. The facts that we (a) have qualia, and (b) claim to have qualia, are related only by coincidence.

I do not think that all forms of epiphenomenalism are ruled out, however. It might be possible to have a principled epiphenomenalism, where the mysterious "dualistic principles" above are replaced by something more concrete, showing that mental states arise in virtue of specific, functionally specifiable physical states. If this were done, then it is possible that the Coherence Test might after all be satisfied: we would *be conscious* and *claim consciousness* in virtue of similar functional accounts. But all that this possibility shows, for now, is that epiphenomenalism would need a great deal of development to be a plausible theory of mind.

Theory 3: Identity Theories, Functional and Physical

I have been rather sympathetic to functionalism in the preceding discussion, but it is not close to a completed theory of mind. It performs well on part (C2) of the Coherence Test. The source of our consciousness-claims is undoubtedly functionally specifiable. True, the flat statement of functionalism gives no insight into just *which* functional properties of the brain are responsible for our consciousness-claims, so there is more work to be done on that point - but we might set that aside as a mere technical difficulty, a matter for cognitive scientists. A more profound difficulty with functionalism is that it does not come close to dealing with (C1). It has often been noted (recently by Searle 1989) that functionalism is not a theory which is motivated by the existence of first-person mentality; rather, conscious mental states are viewed as an *obstacle* for functionalism to deal with (witness all the Absent Qualia and Inverted Spectrum objections). It is possible that functionalism is *compatible* with a correct theory of consciousness, but taken alone it is not that theory.

Functionalism gives very little insight into why consciousness exists. It is one thing to baldly state "subjective experience arises in virtue of our functional organization"; it is another to give an account of *why*. Functionalism, in this way, is not unlike its predecessor, the Identity Theory, which stated "mental states are brain states". The flat statement alone gives us no insight into *how* it could be true. There was, of course, some kernel of truth to the Identity Theory: this was that conscious mental states are indeed *supervenient* on brain states (that is, reproducing a brain state will reproduce the same corresponding mental state). In a similar way, Functionalism contains a kernel of truth: it is very plausible that conscious mental states supervene on certain functional states. By going from the loose notion of "physical state" to the more specific notion of functional state, functionalism has brought us closer to the goal, but it certainly has not achieved it.

So both functionalism and the Identity Theory pass over the key question of just *why* we are conscious in the first place. And of course, the fact that it fails on (C1) implies automatic failure on (C3) - without giving an account of the existence of subjectivity, it cannot give an account of coherence. Nevertheless, I believe that functionalism is a valuable theory, of which many aspects are worth retaining. Something simply needs to be added. The account that I will give below might be characterized as "functionalism plus something".

To satisfy the Coherence Test, and thus be a candidate for the title of "completed theory of mind", a theory must provide three things, on my estimation:

(1) A *metaphysical* account of how subjective experience is possible. This will almost certainly include some new metaphysical construct, over and above physical laws.

(2) A *functional* account of why we think and claim that we are conscious. This will presumably be in the idiom of cognitive science.

(3) An account of functional-metaphysical coherence, which shows how (1) and (2) cohere with each other.

As we have seen, theories that consciousness is biochemical pass the buck on (2) and fail miserably on (3). Epiphenomenalism makes a start on (1), but as it stands has no chance on (3). And functionalism, as currently conceived, does not deal at all with (1), and therefore with (3). I hope that I have demonstrated just how far short currently conceived theories fall of yielding a solution to the Mind-Body Problem. The mode will now switch from pessimism to optimism. The three requirements above sound daunting, but I believe that they might be satisfied. In the next section, I will outline the beginnings of a speculative theory which has a chance of satisfying (1), (2) and (3) above. This theory alone will not be a solution to the Mind-Body Problem, but I hope that it points in the right direction.

Mind, pattern, information

The theory I am going to suggest is an example of what is known as a "double aspect" theory. That is, it holds that first-person mental states and third-person mental states are two different aspects of the same thing. Double aspect theories have been attractive to others (e.g. Nagel 1986), but they have usually foundered because we cannot say exactly what those aspects might be. The simple idea "double aspect theory" explains very little. But before I go on to elaborate on this particular theory, it might be useful to say a couple of words about why double aspect theories are so attractive. In short, this is because double aspect theories are at the same time almost epiphenomenalism and almost identity theories, combining the virtues of both with the vices of neither. The theory I will propose is almost an identity theory (by this term I include both "Brain State" and "Functional State" Identity Theories, though the proposed theory is nearer to the latter than the former), in that it holds that first-person and third-person states are the same thing - but different aspects of the same thing, a crucial difference. (You might say: an identity theory that takes subjective states seriously.) It is almost a version of epiphenomenalism, as we can imagine the subjective aspects "hanging off" the non-subjective aspects, allowing the complete autonomy of the physical - while at the same time allowing that subjective states can be causally efficacious, as they are but another aspect of the objective states. (You might say: an epiphenomenalism where mind can matter.) The account which follows is brief, speculative, poorly-defined and incomplete. Nevertheless, I hope that it offers a glimpse of what might be the correct direction for the understanding of consciousness.

The two different "aspects" that I propose are pattern and information. Wherever they occur, pattern and information occur together. All information is carried by some pattern in the physical world; all patterns carry some information. I speculate that they should be regarded as two different aspects of the same thing, which for want of a better term we may call "pattern/information."

My proposal is that third-person (objectively analyzable) mental events are *patterns* in the brain, while the corresponding subjective (first-person) mental events are *information*. On the above view of pattern/information, this implies that first-person and third-person mental events are indeed two different aspects of the same thing, the underlying pattern/information. (This has the advantage of not expanding our ontology *too far*.)

The idea that third-person mental events can be regarded as patterns should not be too controversial. A commitment very much like this is already made by functionalism, though it is not always emphasized. When we give a functional account of a system, we are necessarily abstracting away from many superficial details, and highlighting particular patterns in the physical substrate. When the functionalist says that a mental state is a functional state, she is committing herself to the notion that a mental state is an *abstraction* - and every abstraction is an abstraction of a pattern. (Of course, not every pattern is a functional pattern, but these are a very important subset.)

It is rarely made clear just what ontological claims a functionalist would wish to make, though. Are they committed to patterns as part of their natural ontology? They might not necessarily have to take this step, as if we are considering mental events only from the third-person viewpoint, then it is not clear that we have to admit them into our ontology anymore than we have to admit centers of gravity - both might be "convenient fictions", in Dennett's terminology. But I am prepared to bite the bullet, and reify patterns. Given that we believe that the mind and conscious experience are part of the basic ontology of things,

then some concession is necessarily -- we have to posit some metaphysical construction of the kind mentioned in the previous section. Admitting patterns into our ontology is a surprisingly painless way of doing this - and once it is done, we get minds for free.

Of course, we have done a little more than reify pattern: we have reified pattern/information. I claim that this is a reasonable step, given that they always go together. Put it this way: if there is any chance that a double aspect theory might be the right theory - and as we have seen and will see, there is very much that is attractive about them - then it is hard to imagine two more natural candidates for the "aspects" than pattern and information. To say that "mental events arise from a double aspect" is to say very little, unless we want to admit the mental as a primitive member of our ontology. Instead, by positing pattern/information, we have a much more natural primitive in our ontology, from which the double-aspect nature of mentality is inherited.

Once we have made this posit, then conscious mentality falls out. Third-person mental events are patterns in the brain: the corresponding conscious mental events are the information that these patterns carry. *Qualia are just information.* (A nice way of putting this is "information is what pattern is like from the inside". I am not quite sure what this means, but it certainly sounds good.) Anyway, conscious mentality arises from the one big pattern that I *am*. That pattern, at any given time, carries a lot of information - that information is my conscious experience. Incidentally, given any physical substrate there are many different ways of abstracting patterns from it. In my brain, there will be some patterns that do not correspond to mental states at all, and some which correspond to unconscious states. This is not a problem for the theory; there will be more on this later.

Why we think we are conscious

The last few paragraphs may strike the reader as wanton ontological extravagance. But apart from any inherent elegance the theory might have, it is also motivated by the fact that it can pass the Coherence Test. In fact, the above is at least in part motivated by an functional account of why we think we are conscious, and why consciousness seems strange. Logically, the functional account ought to come first, but it makes sense to present the metaphysical theory up front. But what follows can be read independently of what went before. This will be a purely functional account, in the tradition of cognitive science. So at least temporarily, all the metaphysical baggage may be thrown away.

Very briefly, here is what I believe to be the correct account of why we think we are conscious, and why it seems like a mystery. The basic notion is that of *pattern processing*. This is one of the things that the brain does best. It can take raw physical data, usually from the environment but even from the brain itself, and extract patterns from these. In particular, it can discriminate on the basis of patterns. The original patterns are in the environment, but they are transformed on their path through neural circuits, until they are represented as quite different patterns in the cerebral cortex. This process can also be represented as *information flow* (not surprisingly), from the environment into the brain. The key point is that once the information flow has reached the central processing portions for the brain, further brain function is not sensitive to the original raw data, but only to the pattern (to the information!) which is embodied in the neural structure.

Consider color perception, for instance. Originally, a spectral envelope of light-wavelengths impinges upon our eyes. Immediately, some distinctions are collapsed, and some pattern is processed. Three different kinds of cones abstract out information about how much light is present in various overlapping wavelength-ranges. This information travels down the optic nerve (as a physical pattern, of course), where it gets further transformed by neural processing into an abstraction about how much intensity is present on what we call the red-green, yellow-blue, and achromatic scales. What happens after this is poorly-understood, but there is no doubt that by the time the central processing region is reached, the pattern is very much transformed, and the information that remains is only an abstraction of certain aspects of the original data.

Anyway, here is why color perception seems strange. In terms of further processing, we are sensitive not to the original data, not even directly to the physical structure of the neural system, but only to the *patterns* which the system embodies, to the information it contains. It is a matter of *access*. When our linguistic system (to be homuncular about things) wants to make verbal reports, it cannot get access to the original data; it does not even have direct access to neural structure. It is sensitive only to pattern. Thus, we *know* that we can make distinctions between certain wavelength distributions, but we do not know how we do it. We've lost access to the original wavelengths - we certainly cannot say "yes, that patch is saturated with 500-600 nm reflections". And we do not have access to our neural structure, so we cannot say "yes, that's a 50 Hz spiking frequency". It is a distinction that we are able to make, but only on the basis of *pattern*. We can merely say "Yes, that looks different from that." When asked "How are they different?", all we can say is "Well, that one's *red*, and that one's *green*". We have access to nothing more - we can simply make raw distinctions based on pattern - and it seems very strange.

So this is why conscious experience seems strange. We are able to make distinctions, but we have direct access neither to the sources of those distinctions, or to how we make the distinctions. The distinctions are based purely on the information that is processed. Incidentally, it seems that the more abstract the information-processing - that is, the more that distinctions are collapsed, and information recoded - the stranger the conscious experience seems. Shape- perception, for instance, strikes us as relatively non-strange; the visual system is extremely good at preserving shape information through its neural pathways. Color and taste are strange indeed, and the processing of both seems to involve a considerable amount of recoding.

The story for "internal perception" is exactly the same. When we reflect on our thoughts, information makes its way from one part of the brain to another, and perhaps eventually to our speech center. It is to only certain abstract features of brain structure that the process is sensitive. (One might imagine that if somehow reflection could be sensitive to every last detail of brain structure, it would seem very different.) Again, we can perceive only via pattern, via information. The brute, seemingly non-concrete distinctions thus entailed are extremely difficult for us to understand, and to articulate. That is why consciousness seems strange, and that is why the debate over the Mind-Body Problem has raged for thousands of years.

The above account is far too brief, and is almost certainly wrong in many important aspects.

Nevertheless, I believe there is a kernel of truth in there. Even if the above account needs to be thoroughly revised, I think this fact will remain: the facts that we think we are conscious, we claim to be conscious, and consciousness seems strange all hold true in virtue of the fact that the brain is a sophisticated pattern-processor (is a sophisticated information-processor).

Further, this account should not surprise us. Things *have* to be this way. Any being which "perceives" the world must do so in virtue of pattern/information processing. Once the processing has reached the heart of the cognitive system, further processing can *only* be sensitive to the information distinctions embodied there. When the being tries to articulate the nature of its perception, it will be reduced to talk of "qualitative distinctions". Similarly, if a being has any reflective access to its internal processing, it will only be in terms of high-level patterns. We should *expect* such beings to be baffled by this "consciousness". The Surprise Principle is vindicated: Consciousness should surprise us, claims about consciousness should not.

The above account, of course, did not rely on any *ontological* commitment to pattern and information - it merely used them as "convenient fictions." This was a completely functional account. If I were a third-person aficionado, who believed that the Mind-Body Problem is only a pseudo-problem, I might stop now, and say "See! Consciousness is just an illusion." But I am not, and this account was intended not to explain away consciousness but to cohere with it.

So now, we may run the Coherence Test. We *claim* that we are conscious in virtue of the brain's ability as a pattern-processor (as an information-processor). We in fact *are* conscious in virtue of the patterns (information) embodied by the brain. These explanations seem to cohere rather well. Of course, there is no pattern-processing without patterns. It is precisely the same patterns which are at the crux of the pattern processing (those at the heart of our processing system, from which all distinctions are made), which are also the patterns which *are* our mental states. (I occasionally use "pattern" as a shorthand for "pattern/ information". No great harm is done.) One could not ask for better coherence than this. The reason we talk about mental states is that our processing is sensitive to precisely those patterns which are our mental states.

Another, slightly over-simplified way to put this, is: (1) The brain perceives itself as pattern. (2) The mind (that we experience) is pattern. Therefore (3) The brain perceives itself as the mind.

Patterns in pattern-processors

On this account, there are two criteria for being a conscious entity: a metaphysical (first-person) and a functional (third-person) criterion. The metaphysical criterion is that one must be a pattern. All patterns exist, presumably. It would seem strange to reify some patterns but not others. But not all patterns are conscious. To be a conscious pattern, one must be part of the right kind of pattern processor, bearing an appropriate relation to it.

An interesting and difficult question concerns the status of patterns which are not part of pattern

processors. Such patterns may still carry information; is there anything it is like to *be* them? My answer is a tentative "yes". But it would not resemble what it is like to be a human being, for such patterns are not *conscious*, in the appropriate functional sense. They are not parts of pattern-processors, so they cannot be aware. This incidentally seems to demonstrate that we should separate the purely first-person notion of "be-ability" from the partially third-person notion of "consciousness". Consciousness itself is dependent on certain functional criteria, but it seems implausible that such third-person criteria should impose restrictions on metaphysical be-ability. Being the number 5 might be like *something*; if only like being asleep, without the excitement of dreaming.

On the other hand, there are pattern processors which are not embodied in human brains. Might such processors (or the patterns therein) have subjective experience? A connectionist network, for instance, is a pattern processor *par excellence*. A typical feed-forward network might have served as a perfect simplified example in the above account of the kind of pattern/information-processing that goes on in the human brain. Networks are sensitive to certain patterns in environmental data; they recode these patterns as they pass down the information-processing chain, until output is sensitive only to a certain kind of pattern in the inputs. Do connectionist networks have qualia? My answer is "yes!" This may seem very counter-intuitive at first, but I believe that upon reflection it becomes plausible. There is no principled distinction between the kind of pattern-processing performed by such a network and that performed by a brain, except one of complexity. And there is no reason to believe that complexity should make such a great difference to the existence of qualia. Certainly human qualia will be more complex and interesting, reflecting the more complex processing; but then, even considering our processing of "redness", there is no evidence that much more complex processing goes into this than goes into a typical large connectionist network. Pattern-processing leads to qualia in humans, and there seems no reason to deny that pattern-processing leads to qualia in networks also. (Of course networks cannot as yet *reflect* on the fact that they have qualia, let alone *talk* about it, but this is not the point.)

Then there is that old favorite, the thermostat. Do thermostats have qualia? It is not entirely clear that thermostats do the right kind of pattern-processing, but if they do, their qualia are remarkably simple. For thermostats process all physical inputs down to three states: too hot, too cold, and just right. It is not clear just what three-valued qualia would be like. But if we are ever to understand qualia, this is a good test case.

Connectionist networks help illustrate another point: that patterns may supervene on functional descriptions. The relevant pattern in networks are *patterns of activation* over a number of units, where units and their activations are of course functionally specifiable. It may not be the case that *all* patterns supervene on functional descriptions (that is, are specifiable relative to functional descriptions, and duplicating the functional system in another substrate allows duplicating the pattern), but it seems plausible that all patterns that represent first-person mental states should. The plausibility lies in the fact that patterns need to be *processed* for us to be aware of them, and processing is a functional notion. The current account is thus compatible with functionalism of a certain variety. You might say that it takes functionalism as a starting point, and adds what is necessary to deal with the problems of consciousness.[*]

*[[No theory of mind worth its salt would pass up the chance to attack Searle's "Chinese Room" problem. For our theory, the explanation is straightforward. When we have a homunculus manipulating symbols on paper, there are two quite distinct sets of patterns: patterns carried by the head of the homunculus, and patterns carried by the complex system of symbols on pattern. The patterns in the paper may well support their own mind. Even if we assume the homunculus could internalize the rules (which would require vastly more memory than any human has), then there would still be two sets of quite distinct patterns, both of which are present in the homunculus's head. It is a fundamental fact that there can be many different patterns present in a given substrate. Thus, in a single head, there might be two quite distinct phenomenologies, without any overlapping first-person mental states.]]

Loose ends

Moving away from the more cognitive aspects to the more metaphysical aspects, we are yet to explicitly answer our original question (4c): how can consciousness have any bearing on what we say about consciousness? On the present account, this is easy. Conscious experience is identified with the "information" aspect of certain pattern/information states. This information can certainly make a difference. It is just another aspect of the pattern, and there is no question that the pattern plays a causal role. Changing the information changes the pattern, and changing the pattern changes many consequences. In fact, as we have seen, our claims about consciousness reflect precisely that pattern/information in which our mental states consist. Consciousness is thus causally efficacious. At the same time, it is quite possible to analyze third-person mental states and thus behavior without ever needing to invoke consciousness, if we so desire. It is possible to cast everything in terms of pattern, without invoking information.

This answer corresponds to none of the answers (A), (B) or (C) given originally. Our options were in fact not quite so limited as they seemed. This again shows the advantages of a double aspect theory - an identity theory which takes consciousness seriously.

It might be noted that this is not quite a double aspect theory in the traditional sense, where the two aspects are the physical and the mental. This theory holds instead that the two aspects are first-person and third-person mental states - both of them mental, though in different senses. We are able to use the resources of theories like functionalism to get us past the first stage, from the physical to third-person mental states, which may be functional states, computational states, or some other patterns. All that is left is Jackendoff's "Mind-Mind" Problem - the bridge from third-person patterns (the "Computational Mind") to first-person mental states (the "Phenomenological Mind"). This is precisely what our posited pattern/information duality achieves.

Finally, it must be glaringly obvious that I have said very little about a very important topic: the relationship between pattern and information. Is this merely a primitive posited relation, or can it somehow be explicated? Some might claim that a given pattern can be *interpreted* as carrying any information you like. I believe that this is using a different sense of the word "information" to mine. The kind of information I have dealt with here is not "information *that*" - it does not need to refer. The kind of information I believe in is intrinsic. In this sense, there is a *natural mapping* from pattern to information.

Exactly what the nature of this mapping is is very difficult to say. This is why the current theory is not yet a solution to the Mind-Body Problem. It may be the case that the pattern/information relation is a brute fact, in much the same way that the laws of physics are brute facts. We have a certain amount of knowledge about their nomological connection, via our subjective experience. This may or may not be enough to understand all the facts perfectly. If it is not, then it may be the case that we can never know precisely what it is like to be an alien being (or a connectionist network), as we will not understand precisely the nature of the information that their patterns carry. The outlook is not necessarily so pessimistic, though, and perhaps a better analysis of the nature of pattern and information will tell us everything there is to know. (If pressed for an informal opinion, I would guess that at least nine tenths of the pattern/information relationship will be understandable analytically, with the last tenth up for grabs. If we are not able to understand *precisely* what it is like to be a bat, we will be able to come very close.)

This account has not removed the Mind-Body Problem. Nevertheless, my minimal hope is that it shows what the beginnings of a solution might look like. My maximal hope is that it has removed much of the confusion surrounding the problem, and localized the mystery at one key, primitive locus: the relationship between pattern and information.

References

Block, N. 1980. Are absent qualia impossible? *Philosophical Review* 89:257-74.

Churchland, P.M. 1985. Reduction, qualia and the direct introspection of brain states. *Journal of Philosophy* 82:8-28.

Dennett, D.C. 1982. How to study human consciousness empirically, or, nothing comes to mind. *Synthese* 53:159-80.

Dennett, D.C. 1978. Toward a cognitive theory of consciousness. *Minnesota Studies in the Philosophy of Science*, Vol. 9. Reprinted in *Brainstorms* (MIT Press, 1978).

Dennett, D.C. 1988. Quining qualia. In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

Elitzur, A. 1989. Consciousness and the incompleteness of the physical explanation of behavior. *Journal Of Mind and Behavior* 10:1-20.

Foss, J. 1989. On the logic of what it is like to be a conscious subject. *Australasian Journal of Philosophy* 67: 205-220.

Harnad, S.R. 1989. Minds, machines and Searle. *Journal of Theoretical and Experimental Artificial Intelligence* 1:5-26.

- Hofstadter, D.R. 1979. *Gödel, Escher, Bach*. New York: Basic Books.
- Horgan, T. 1984. Functionalism, qualia, and the inverted spectrum. *Philosophy and Phenomenological Research* 44: 453-469.
- Jackendoff, R.S. 1987. *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.
- Jackson, F.C. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32: 127-136.
- Johnson-Laird, P.N. 1983. A computational analysis of consciousness. *Cognition and Brain Theory* 6: 499-508.
- McGinn, C. 1989. Can We Solve The Mind-Body Problem? *Mind* 98:349-366.
- Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 83:435-450.
- Nagel, T. 1986. *The View From Nowhere*. Oxford: Oxford University Press.
- Parfit, D.A. 1984. *Reasons and Persons*. Oxford:Oxford University Press.
- Rumelhart, D.E., McClelland, J.R., and the PDP Research Group 1986. *Parallel Distributed Processing*. Cambridge, MA: MIT Press.
- Searle, J.R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-424.
- Searle, J.R. 1989. Consciousness, unconsciousness, and intentionality. *Philosophical Topics* 17:193-209.
- Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:271-315.

On the Search for the Neural Correlate of Consciousness[*]

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

*[[This paper appears in *Toward a Science of Consciousness II: The Second Tucson Discussions and Debates* (S. Hameroff, A. Kaszniak, and A. Scott, eds), published with MIT Press in 1998. It is a transcript of my talk at the second Tucson conference in April 1996, lightly edited to include the contents of overheads and to exclude some diversions with a consciousness meter. A more in-depth argument for some of the claims in this paper can be found in Chapter 6 of my book *The Conscious Mind* (Chalmers, 1996).]]

I'm going to talk about one aspect of the role that neuroscience plays in the search for a theory of consciousness. Whether or not neuroscience can solve all the problems of consciousness singlehandedly, there is no question that it has a major role to play. We've seen at this conference that there's a vast amount of progress in neurobiological research, and that much of it is clearly bearing on the problems of consciousness. But the conceptual foundations of this sort of research are only beginning to be laid. So I will look at some of the things that are going on from a philosopher's perspective and will see if there's anything helpful to say about these foundations.

We've all been hearing a lot about the "neural correlate of consciousness". This phrase is intended to refer to the neural system or systems primarily associated with conscious experience. I gather that the catchword of the day is "NCC". We all have an NCC inside our head, we just have to find out what it is. In recent years there have been quite a few proposals about the identity of the NCC. One of the most famous proposals is Crick and Koch's suggestion concerning 40-hertz oscillations. That proposal has since faded away a little but there are all sorts of other suggestions out there. It's almost got to a point where it's reminiscent of particle physics, where they have something like 236 particles and people talk about the "particle zoo". In the study of consciousness, one might talk about the "neural correlate zoo". There have also been a number of related proposals about what we might call the "cognitive correlate of consciousness" (CCC?).

A small list of suggestions that have been put forward might include:

- 40-hertz oscillations in the cerebral cortex (Crick and Koch 1990)
- Intralaminar nucleus in the thalamus (Bogen 1995)
- Re-entrant loops in thalamocortical systems (Edelman 1989)

- 40-hertz rhythmic activity in thalamocortical systems (Llinas et al 1994)
- Nucleus reticularis (Taylor and Alavi 1995)
- Extended reticular-thalamic activation system (Newman and Baars 1993)
- Anterior cingulate system (Cotterill 1994)
- Neural assemblies bound by NMDA (Flohr 1995)
- Temporally-extended neural activity (Libet 1994)
- Backprojections to lower cortical areas (Cauller and Kulics 1991)
- Neurons in extrastriate visual cortex projecting to prefrontal areas (Crick and Koch 1995)
- Neural activity in area V5/MT (Tootell et al 1995)
- Certain neurons in the superior temporal sulcus (Logothetis and Schall 1989)
- Neuronal gestalts in an epicenter (Greenfield 1995)
- Outputs of a comparator system in the hippocampus (Gray 1995)
- Quantum coherence in microtubules (Hameroff 1994)
- Global workspace (Baars 1988)
- Activated semantic memories (Hardcastle 1995)
- High-quality representations (Farah 1994)
- Selector inputs to action systems (Shallice 1988)

There are a few intriguing commonalities among the proposals on this list. A number of them give a central role to interactions between the thalamus and the cortex, for example. All the same, the sheer number and diversity of the proposals can be a little overwhelming. I propose to step back a little and try to make sense of all this activity by asking some foundational questions.

A central question is this: how is it, in fact, that one can search for the neural correlate of consciousness? As we all know, there are problems in measuring consciousness. It's not a directly and straightforwardly observable phenomenon. It would be a lot easier if we had a way of getting at consciousness directly; if we had, for example, a consciousness meter.

If we had a consciousness meter, searching for the NCC would be straightforward. We'd wave the consciousness meter and measure a subject's consciousness directly. At the same time, we'd monitor the underlying brain processes. After a number of trials, we'd say OK, such-and-such brain processes are correlated with experiences of various kinds, so that's the neural correlate of consciousness.

Alas, we don't have a consciousness meter, and there seem to be principled reasons why we can't have one. Consciousness just isn't the sort of thing that can be measured directly. So: What do we do without a consciousness meter? How can the search go forward? How does all this experimental research proceed?

I think the answer is this: we get there through principles of *interpretation*. These are principles by which we interpret physical systems to judge whether or not they have consciousness. We might call these *pre-experimental bridging principles*. These are the criteria that we bring to bear in looking at systems to say (a) whether or not they are conscious now, and (b) what information they are conscious of, and what information they are not. We can't reach in directly and grab those experiences and "transpersonalize"

them into our own, so we rely on external criteria instead.

That's a perfectly reasonable thing to do. But in doing this we have to realize that something interesting is going on. These principles of interpretation are not themselves experimentally determined or experimentally tested. In a sense they are pre-experimental assumptions. Experimental research gives us a lot of information about processing; then we bring in the bridging principles to interpret the experimental results, whatever those results may be. They are the principles by which we make *inferences* from facts about processing to facts about consciousness, so they are conceptually prior to the experiments themselves. We can't actually refine them experimentally (except perhaps through first-person experimentation!), because we don't have any independent access to the independent variable. Instead, these principles will be based on some combination of (a) conceptual judgments about what counts as a conscious process and (b) information gleaned from our first-person perspective on our own consciousness.

I think we are all stuck in this boat. The point applies whether one is a reductionist or an anti-reductionist about consciousness. A hard-line reductionist might put some of these points slightly differently, but either way, the experimental work is going to require pre-experimental reasoning to determine the criteria for ascription of consciousness. Of course such principles are usually left implicit in empirical research. We don't usually see papers saying "Here is the bridging principle, here are the data, and here is what follows." But it's useful to make them explicit. The very presence of these principles has some strong and interesting consequences in the search for the NCC.

In a sense, in relying on these principles we are taking a leap into the epistemological unknown. Because we don't measure consciousness directly, we have to make something of a leap of faith. It may not be a big leap, but nevertheless it suggests that everyone doing this sort of work is engaged in philosophical reasoning. Of course one can always choose to stay on solid ground, talking about the empirical results in a neutral way; but the price of doing so is that one gains no particular insight into consciousness. Conversely, as soon as we draw any conclusions about consciousness, we have gone beyond the information given, so we need to pay careful attention to the reasoning involved.

So what are these principles of interpretation? The first and by far the most prevalent such principle is a very straightforward one: it's a principle of verbal report. When someone says "Yes, I see that table now", we infer that they are conscious of the table. When someone says "Yes, I see red now", we infer that they are having an experience of red. Of course one might always say "How do you know?" -- a philosopher might suggest that we may be faced with a fully functioning zombie - but in fact most of us don't believe that the people around us are zombies, and in practice we are quite prepared to rely on this principle. As pre-experimental assumptions go, this is a relatively "safe" one - it doesn't require a huge leap of faith - and it is very widely used.

So the principle here is that when information is verbally reported, it is conscious. One can extend this slightly, as no-one believes that an *actual* verbal report is required for consciousness; we are conscious of much more than we report on any given occasion. So an extended principle might say that when

information is directly *available* for verbal report, it is conscious.

Experimental researchers don't rely only on these principles of verbal report and reportability. These principles can be somewhat limiting when we want to do broader experiments. In particular, we don't want to just restrict our studies of consciousness to subjects that have language. In fact just this morning we saw a beautiful example of research on consciousness in language-free creatures. I'm referring to the work of Nikos Logothetis and his colleagues (e.g. Logothetis & Schall 1989; Leopold & Logothetis 1996). This work uses experiments on binocular rivalry in monkeys to draw conclusions about the neural processes associated with consciousness. How do Logothetis *et al* manage to draw conclusions about a monkey's consciousness without getting any verbal reports? What they do is rely on a monkey's pressing bars: if a monkey can be made to press a bar in an appropriate way in response to a stimulus, we'll say that that stimulus was consciously perceived.

The criterion at play seems to be require that the information be available for an arbitrary response. If it turned out that the monkey could press a bar in response to a red light but couldn't do anything else, we would be tempted to say that it wasn't a case of consciousness at all, but some sort of subconscious connection. If on the other hand we find information that is available for response in all sorts of different ways, then we'll say that it is conscious. Actually Logothetis and his colleagues also use some subtler reasoning about similarities with binocular rivalry in humans to buttress the claim that the monkey is having the relevant conscious experience, but it is clearly the response that carries the most weight.

The underlying general principle is something like this: When information is *directly available for global control* in a cognitive system, then it is conscious. If information is available for response in many different motor modalities, we will say that it is conscious, at least in a range of relatively familiar systems such as humans and primates and so on. This principle squares well with the previous principle in cases where the capacity for verbal report is present: availability for verbal report and availability for global control seem to go together in such cases (report is one of the key aspects of control, after all, and it is rare to find information that is reportable but not available more widely). But this principle is also applicable more widely.

A correlation between consciousness and global availability (for short) seems to fit the first-person evidence - the evidence gleaned from our own conscious experience - quite well. When information is present in my consciousness, it is generally reportable, and it can generally be brought to bear in the control of behavior in all sorts of different ways. I can talk about it, I can point in the general direction of a stimulus, I can press bars, and so on. Conversely, when we find information that is directly available in this way for report and other aspects of control, it is generally conscious information. I think one can bear this out by consideration of cases.

There are some interesting puzzle cases to consider, such as the case of blindsight, where one has *some* kind of availability for control but arguably no conscious experience. Those cases might best be handled by invoking the directness criterion: insofar as the information here is available for report and other control processes at all, it is available only indirectly, by comparison to the direct and automatic

availability in standard cases. One might also stipulate that it is availability for *voluntary* control that is relevant, to deal with certain cases of involuntary unconscious response, although that is a complex issue. I discuss a number of puzzle cases in more detail elsewhere (Chalmers 1996, forthcoming), where I also give a much more detailed defence of the idea that something like global availability is the key pre-empirical criterion for the ascription of consciousness.

But this remains at best a first-order approximation of the functional criteria that come into play. I'm less concerned today to get all the fine details right than to work with the idea that some such functional criterion is required and indeed is implicit in all the empirical research on the neural correlate of consciousness. If you disagree with the criterion I've suggested here - presumably because you can think of counterexamples - you may want to use those counterexamples to refine it or to come up with a better criterion of your own. But the point I want to focus on here is that in the very act of experimentally distinguishing conscious from unconscious processes, *some* such criterion is always at play.

So the question I want to ask is: if *something* like this is right, then what follows? That is, if some such bridging principles are implicit in the methodology of the search for the NCC, then what are the consequences? I will use global availability as my central functional criterion in the discussion that follows, but many of the points should generalize.

The first thing one can do is produce what philosophers might call a *rational reconstruction* of the search for the neural correlate of consciousness. With a rational reconstruction we can say, maybe things don't work exactly like this in practice, but the rational underpinnings of the process have something like this form. That is, if one were to try to *justify* the conclusions one has reached as well as one can, one's justification would follow the shape of the rational reconstruction. In this case, a rational reconstruction might look something like this:

(1) Consciousness $\leftrightarrow$ global availability (bridging principle)

(2) Global availability $\leftrightarrow$ neural process N (empirical work)

so

(3) Consciousness $\leftrightarrow$ neural process N (conclusion).

According to this reconstruction, one implicitly embraces some sort of pre-experimental bridging principle that one finds plausible on independent grounds, such as conceptual or phenomenological grounds. Then one does the empirical research. Instead of measuring consciousness directly, we detect the functional property. One sees that when this functional property (e.g. global availability) is present, it is correlated with a certain neural process (e.g. 40-hertz oscillations). Combining the pre-empirical premise and the empirical result, we arrive at the conclusion that this neural process is a candidate for the NCC.

Of course it doesn't work nearly so simply in practice. The two stages are very intertwined; our pre-experimental principles may themselves be refined as experimental research goes along. Nevertheless I think one can make a separation, at least at the rational level, into pre-empirical and experimental components, for the sake of analysis. So with this sort of rational reconstruction in hand, what sort of conclusions follow? There are about six consequences that I want to draw out here.

(1) The first conclusion is a characterization of the neural correlates of consciousness. If the NCC is arrived at through this sort of methodology, then whatever it turns out to be, it will be a *mechanism of global availability*. The presence of the NCC wherever global availability is present suggests that it is a mechanism that *subserves* the process of global availability in the brain. The only alternative that we have to worry about is that it might be a *symptom* rather than a *mechanism* of global availability; but that possibility ought to be addressable in principle by dissociation studies, by lesioning, and so on. If a process is a mere symptom of availability, we ought to be able to empirically dissociate it from the process of global availability while leaving the latter intact. The resulting data would suggest to us that consciousness can be present even when the neural process in question is not, thus indicating that it wasn't a perfect correlate of consciousness after all.

(A related line of reasoning supports the idea that a true NCC must be a mechanism of *direct* availability for global control. Mechanisms of indirect availability will in principle be dissociable from the empirical evidence for consciousness, for example by directly stimulating the mechanisms of direct availability. The indirect mechanisms will be "screened off" by the direct mechanisms in much the same way as the retina is screened off as an NCC by the visual cortex.)

In fact, if one looks at the various proposals that are out there, this template seems to fit them pretty well. For example, the 40-hertz oscillations discussed by Crick and Koch were put forward precisely because of the role they might play in binding and integrating information into working memory, and working memory is of course a central mechanism whereby information is made available for global control in a cognitive system. Similarly, it is plausible that Libet's extended neural activity is relevant precisely because the temporal extendedness of activity is what gives certain information the capacity to dominate later processes that lead to control. Baars' global workspace is a particularly explicit proposal of a mechanism in this direction; it is put forward explicitly as a mechanism whereby information can be globally disseminated. All of these mechanisms and many of the others seem to be candidates for mechanisms of global availability in the brain.

(2) This reconstruction suggests that a full story about the neural processes associated with consciousness will to do two things. Firstly, it will *explain* global availability in the brain. Once we know all about the relevant neural processes, we will know precisely how information is made directly available for global control in the brain, and this will be an explanation in the full sense. Global availability is a functional property, and as always the problem of explaining the performance of a function is a problem to which

mechanistic explanation is well-suited. So we can be confident that in a century or two global availability will be straightforwardly explained. Secondly, this explanation of availability will do something else: it will isolate the processes that *underlie* consciousness itself. If the bridging principle is granted, then mechanisms of availability will automatically be correlates of phenomenology in the full sense.

Now, I don't think this gives us a full *explanation* of consciousness. One can always raise the question of why it is that these processes of availability should give rise to consciousness in the first place. As yet we have no explanation of why this is, and it may well be that the full details concerning the processes of availability still won't answer this question. Certainly, nothing in the standard methodology I have outlined answers the question; that methodology *assumes* a relation between availability and consciousness, and therefore does nothing to *explain* it. The relationship between the two is instead taken as something of a primitive. So the hard problem still remains. But who knows: somewhere along the line we may be led to the relevant insights that show why the link is there, and the hard problem may then be solved. In any case, whether or not we have solved the hard problem, we may nevertheless have isolated the *basis* of consciousness in the brain. We just have to keep in mind the distinction between correlation and explanation.

(3) Given this paradigm, it is likely that there are going to be many different neural correlates of consciousness. I take it that this is not going to surprise many people; but the rational reconstruction gives us a way of seeing just why such a multiplicity of correlates should exist. There will be many neural correlates of consciousness because there may well be many different mechanisms of global availability. There will be mechanisms of availability in different modalities: the mechanisms of visual availability may be quite different from the mechanisms of auditory availability, for example. (Of course they *may* be the same, in that we could find a later area that integrates and disseminates all this information, but that's an open question.) There will also be mechanisms at different stages of the processing path whereby information is made globally available: early mechanisms and later ones. So these may all be candidates for the NCC. And there will be mechanisms at many different levels of description: for example, 40-hertz oscillations may well be redescribed as high-quality representations, or as part of a global workspace, at a different level of description. So it may turn out that a number of the animals in the zoo, so to speak, can co-exist, because they are compatible in one of these ways.

I won't speculate much further on just what the neural correlates of consciousness *are*. No doubt some of the ideas in the initial list will prove to be entirely off-track, while some of the others will prove closer to the mark. As we philosophers like to say, humbly, that's an empirical question. But I hope the conceptual issues are becoming clearer.

(4) This way of thinking about things allows one to make sense of a idea that is sometimes floated: that of a *consciousness module*. Sometimes this notion is disparaged; sometimes it is embraced. But this picture of the methodology in the search for an NCC suggests that it is at least possible that there could

turn out to be such a module. What would it take? It would require that there turns out to be some sort of functionally localizable, internally integrated area, through which all global availability runs. It needn't be anatomically localizable, but to qualify as a module it would need to be localizable in some broader sense. For example, the parts of the module would have to have high-bandwidth communication among themselves, compared to the relatively low-bandwidth communication that they have with other areas. Such a thing *could* turn out to exist. It doesn't strike me as especially *likely* that things will turn out this way; it seems just as likely that there will be multiple independent mechanisms of global availability in the brain, scattered around without any special degree of mutual integration. If that's so, we will likely say that there doesn't turn out to be a consciousness module after all. But that's another one of those empirical questions.

If something like this does turn out to exist in the brain, it would resemble Baars' conception of a global workspace: a functional area responsible for the integration of information in the brain and for its dissemination to multiple nonconscious specialized processes. In fact I should acknowledge that many of the ideas I'm putting forward here are compatible with things that Baars has been saying for years about the role of global availability in the study of consciousness. Indeed, this way of looking at things suggests that some of his ideas are almost forced on one by the methodology. The special epistemological role of global availability helps explain why the idea of a global workspace provides a useful way of thinking about almost any empirical proposal about consciousness. If NCC's are identified as such precisely because of their role in global control, then at least on a first approximation, we should expect the global workspace idea to be a natural fit.

(5) We can also apply this picture to a question that has been discussed frequently at this conference: are the neural correlates of *visual* consciousness to be found in V1, in the extrastriate visual cortex, or elsewhere? If our picture of the methodology is correct, then the answer will presumably depend on which visual area is most directly implicated in global availability.

Crick and Koch have suggested that the visual NCC is not to be found within V1, as V1 does not contain neurons that project to the prefrontal cortex. This reasoning has been criticized by Ned Block for conflating access consciousness and phenomenal consciousness (see Block, this volume); but interestingly, the picture I have developed suggests that it may be good reasoning. The prefrontal cortex is known to be associated with control processes; so *if* a given area in the visual cortex projects to prefrontal areas, then it may well be a mechanism of direct availability. And if it does not project in this way, it is less likely to be such a mechanism; at best it might be *indirectly* associated with global availability. Of course there is still plenty of room to raise questions about the empirical details. But the broader point is that for the sort of reasons discussed in (2) above, it is likely that the neural processes involved in *explaining* access consciousness will simultaneously be involved in a story about the *basis* of phenomenal consciousness. If something like this is implicit in their reasoning, Crick and Koch might escape the charge of conflation. Of course the reasoning does depend on these somewhat shaky bridging principles, but then all work on the neural correlates of consciousness must appeal to such principles somewhere, so this can't be held against Crick and Koch in particular.

(6) Sometimes the neural correlate of consciousness is conceived of as the Holy Grail for a theory of consciousness. It will make everything fall into place. For example, once we discover the NCC, then we'll have a definitive test for consciousness, enabling us to discover consciousness wherever it arises. That is, we might use the neural correlate itself as a sort of consciousness meter. If a system has 40-hertz oscillations (say), then it is conscious; if it has none, then it is not conscious. Or if a thalamocortical system turns out to be the NCC, then a system without that system is unlikely to be conscious. This sort of reasoning is not usually put quite so baldly as this, but I think one finds some version of it quite frequently.

This reasoning can be tempting, but one should not succumb to the temptation. Given the very methodology which comes into play here, there is no way to definitely establish a given NCC as an independent test for consciousness. The primary criterion for consciousness will always remain the functional property we started with: global availability, or verbal report, or whatever. That's how we discovered the correlations in the first place. 40-hertz oscillations (or whatever) are relevant *only* because of the role they play in satisfying this criterion. True, in cases where we know that this association between the NCC and the functional property is present, the NCC might itself function as a sort of "signature" of consciousness; but once we dissociate the NCC from the functional property, all bets are off. To take an extreme example, if we have 40-hertz oscillations in a test tube, that almost certainly won't yield consciousness. But the point applies equally in less extreme cases. Because it was the bridging principles that gave us all the traction in the search for an NCC in the first place, it's not clear that anything follows in cases where the functional criterion is thrown away. So there's no free lunch here: one can't get something for nothing.

Once one recognizes the central role that pre-experimental assumptions play in the search for the NCC, one realizes that there are some limitations on just what we can expect this search to tell us. Still, whether or not the NCC is the Holy Grail, I hope that I have said enough to make it clear that the quest for it is likely to enhance our understanding considerably. And I hope to have convinced you that there are important ways in which philosophy and neuroscience can come together to help clarify some of the deep problems involved in the study of consciousness.

References

Baars, B.J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bogen, J.E. 1995. On the neurophysiology of consciousness, parts I and II. *Consciousness and Cognition*, 4:52-62 & 4:137-58.

Cauller, L.J. & Kulics, A.T. 1991. The neural basis of the behaviorally relevant N1 component of the somatosensory evoked potential in awake monkeys: Evidence that backward cortical projections signal

conscious touch sensation. *Experimental Brain Research* 84:607-619.

Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D.J. (forthcoming). Availability: the cognitive basis of experience? *Behavioral and Brain Sciences*. Also in N. Block, O. Flanagan, & G. Güzeldere (eds) *The Nature of Consciousness* (MIT Press, 1997).

Cotterill, R. 1994. On the unity of conscious experience. *Journal of Consciousness Studies* 2:290-311.

Crick, F. and Koch, C. 1990. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2: 263-275.

Crick, F. & Koch, C. 1995. Are we aware of neural activity in primary visual cortex? *Nature* 375: 121-23.

Edelman, G.M. 1989. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

Farah, M.J. 1994. Visual perception and visual awareness after brain damage: A tutorial overview. In (C. Umiltà and M. Moscovitch, eds.) *Consciousness and Unconscious Information Processing: Attention and Performance* 15. MIT Press.

Flohr, H. 1995. Sensations and brain processes. *Behavioral Brain Research* 71:157-61.

Gray, J.A. 1995. The contents of consciousness: A neuropsychological conjecture. *Behavioral and Brain Sciences* 18:659-722.

Greenfield, S. 1995. *Journey to the Centers of the Mind*. W.H. Freeman.

Hameroff, S.R. 1994. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1:91-118.

Hardcastle, V.G. 1996. *Locating Consciousness*. Philadelphia: John Benjamins.

Jackendoff, R. 1987. *Consciousness and the Computational Mind*. MIT Press.

Leopold, D.A. & Logothetis, N.K. 1996. Activity-changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* 379: 549-553.

- Libet, B. 1993. The neural time factor in conscious and unconscious events. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). New York: Wiley.
- Llinas, R.R., Ribary, U., Joliot, M. & Wang, X.-J. 1994. Content and context in temporal thalamocortical binding. In (G. Buzsaki, R.R. Llinas, & W. Singer, eds.) *Temporal Coding in the Brain*. Berlin: Springer Verlag.
- Logothetis, N. & Schall, J. 1989. Neuronal correlates of subjective visual perception. *Science* 245:761-63.
- Shallice, T. 1988. Information-processing models of consciousness: possibilities and problems. In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science*. Oxford University Press.
- Taylor, J.G. & Alavi, F.N. 1993. Mathematical analysis of a competitive network for attention. In (J.G. Taylor, ed.) *Mathematical Approaches to Neural Networks*. Elsevier.
- Tootell, R.B., Reppas, J.B., Dale, A.M., Look, R.B., Sereno, M.I., Malach, R., Brady, J. & Rosen, B.R. 1995. Visual motion aftereffect in human cortical area MT revealed by functional magnetic resonance imaging. *Nature* 375:139-41.

What is a Neural Correlate of Consciousness?

[David J. Chalmers](#)

**Department of Philosophy
University of Arizona
Tucson, AZ 85721.**

chalmers@arizona.edu

[[This paper was published in *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (T. Metzinger, ed), published with MIT Press in 2000. It was first presented at Association for the Scientific Study of Consciousness conference on "Neural Correlates of Consciousness" in Bremen, June 1998.]

1 Introduction

The search for neural correlates of consciousness (or NCCs) is arguably the cornerstone in the recent resurgence of the science of consciousness. The search poses many difficult empirical problems, but it seems to be tractable in principle, and some ingenious studies in recent years have led to considerable progress. A number of proposals have been put forward concerning the nature and location of neural correlates of consciousness. A few of these include:

- 40-hertz oscillations in the cerebral cortex (Crick and Koch 1990)
- Intralaminar nuclei in the thalamus (Bogen 1995)
- Re-entrant loops in thalamocortical systems (Edelman 1989)
- 40-hertz rhythmic activity in thalamocortical systems (Llinas et al 1994)
- Extended reticular-thalamic activation system (Newman and Baars 1993)
- Neural assemblies bound by NMDA (Flohr 1995)
- Certain neurochemical levels of activation (Hobson 1997)
- Certain neurons in inferior temporal cortex (Sheinberg and Logothetis 1997)
- Neurons in extrastriate visual cortex projecting to prefrontal areas (Crick and Koch 1995)
- Visual processing within the ventral stream (Milner and Goodale 1995)

(A longer list can be found in Chalmers 1998. Review articles on neural correlates of consciousness, especially visual consciousness, can be found in Crick and Koch 1998 and Milner 1995.)

As the full title of this book ("Neural Correlates of Consciousness: Empirical and Conceptual Issues") suggests, all this activity raises a number of difficult conceptual and foundational issues. I can see at least five sorts of foundational questions in the vicinity.

- (1) What do we mean by 'consciousness'?

- (2) What do we mean by 'neural correlate of consciousness'?
- (3) How can we find the neural correlate(s) of consciousness?
- (4) What will a neural correlate of consciousness explain?
- (5) Is consciousness reducible to its neural correlate(s)?

The first two questions here are conceptual questions, the third is an epistemological or methodological question, the fourth is an explanatory question, and the fifth is an ontological question. The first, fourth, and fifth are versions of general questions that philosophers have discussed for a long time (my own view on them is in Chalmers (1995; 1996)). The second and third questions are more specific to the NCC investigation. I have discussed the third question in Chalmers (1998). Here I want to focus on the second question.

What does it mean to be a neural correlate of consciousness? At first glance, the answer might seem to be so obvious that the question is hardly worth asking. An NCC is just a neural state that directly correlates with a conscious state, or which directly generates consciousness, or something like that. One has a simple image: when your NCC is active, perhaps, your consciousness turns on, and in a corresponding way. But a moment's reflection suggests that the idea is not completely straightforward, and that the concept needs some clarification.

Here, I will attempt a little conceptual spadework in clarifying the concept of an NCC. I don't know that this is the deepest problem in the area, but it seems to me that if we are looking for an NCC, it makes sense to get clear on what we are looking for. On the way I will try to make contact with some of the empirical work in the area, and see what concept of NCC is at play in some of the central work in the field. I will also draw out some consequences for the methodology of empirical work in the search. Most of this is intended as a first step rather than a last word. Much of what I say will need to be refined, but I hope at least to draw attention to some interesting issues in the vicinity.

As a first pass, we can use the definition of a neural correlate of consciousness given in the program of the ASSC conference. This says a neural correlate of consciousness is a "specific system in the brain whose activity correlates directly with states of conscious experience". This yields something like the following:

A neural system N is an NCC if the state of N correlates directly with states of consciousness.

There are at least two things to get clear on here. First, what are the relevant "states of consciousness"? Second, what does it mean for a neural state to "correlate directly" with states of consciousness? I'll look into both these things in turn.

2 States of consciousness

I will take it that the states of consciousness we are concerned with here are all states of subjective experience, or equivalently, states of phenomenal consciousness. But what sort of states are relevant? In

the NCC literature, I can see a few different classes of state that are sometimes considered.

(i) *Being conscious*

The first option is that the states in question are just those of being conscious and of not being conscious. The corresponding notion of an NCC will be that of a neural system whose state directly correlates with whether a subject is conscious or not. If the NCC is in a particular state, the subject will be conscious. If the NCC is not in that state, the subject will not be conscious.

This is perhaps the idea that first comes to mind when we think about an NCC. We might think about it as the "neural correlate of creature consciousness", where creature consciousness is the property a creature has when it is conscious, and lacks when it is not conscious.

Although this is an interesting notion, it does not seem to capture the sort of NCC that most work in the area is aimed at. As we'll see, most current work is aimed at something more specific. There are some ideas that can be taken as aiming at least in part at this notion, though. For example, the ideas of Bogen (1995) about the intralaminar nucleus seem to be directed at least in part at this sort of NCC.

Examining current work, it's interesting to note that insofar as there is any consensus at all about the location of this sort of NCC, the dominant view seems to be that it should be in or around the thalamus, or at least that it should involve interactions between the thalamic and cortical systems in a central role. Penfield (1937) argued that "the indispensable substratum of consciousness" lies outside the cerebral cortex, and probably lies in the diencephalon (thalamus, hypothalamus, subthalamus, epithalamus). This theme has been taken up in recent years by Bogen, Newman and Baars (1993), and others.

(ii) *Background state of consciousness*

A related idea is that of the neural correlate of what we might call the *background state* of consciousness. A background state is an overall state of consciousness such as being awake, being asleep, dreaming, being under hypnosis, and so on. Exactly what counts as a background state is not entirely clear, as one can divide things up in a number of ways, and with coarser or finer grains, but presumably the class will include a range of normal and of "altered" states.

We can think of this as a slightly more fine-grained version of the previous idea. Creature consciousness is the most coarse-grained background state of consciousness: it is just the state of being conscious. Background states will usually be more fine-grained than this, but they still will not be defined in terms of specific contents or modalities.

A neural correlate of the background state of consciousness, then, will be a neural system N such that the state of N directly correlates with whether a subject is awake, dreaming, under hypnosis, and so on. If N is in state 1, the subject is awake; if N is in state 2, the subject is dreaming; if N is in state 3, the subject is under hypnosis; and so on.

It may well be that some of the thalamocortical proposals discussed above are intended as, or might be extended into proposals about this sort of NCC. A more direct example is given by Hobson's (1997) ideas about neurochemical levels of activation. Hobson holds that these levels can be grouped into a three-dimensional state-space, and that different regions in this space correspond to different overall states of consciousness: wakefulness, REM sleep, nonREM sleep, and so on. When chemical levels are in a particular region in this space, the subject will be awake; when in another region, the subject will be in REM sleep; and so on. On this reading, one might see the neurochemical system as an NCC of the sort characterized above, with the different regions in state space corresponding to correlates of the various specific background states.

(iii) *Contents of consciousness*

There is much more to consciousness than the mere state of being conscious, or the background state of consciousness. Arguably the most interesting states of consciousness are *specific* states of consciousness: the fine-grained states of subjective experience that one is in at any given time. Such states might include the experience of a particular visual image, of a particular sound pattern, of a detailed stream of conscious thought, and so on. A detailed visual experience, for example, might include the experience of certain shapes and colors in one's environment, of specific arrangements of objects, of various relative distances and depths, and so on.

Specific states like these are most often individuated by their *content*. Most conscious states seem to have some sort of specific representational content, representing the world as being one way or another. Much of the specific nature of a visual experience, for example, can be characterized in terms of content. A visual experience typically represents the world as containing various shapes and colors, as containing certain objects standing in certain spatial relations, and so on. If the experience is veridical, the world will be the way the experience represents it as being. If the experience is an illusion or is otherwise misleading, the world will be other than the experience represents it as being. But either way, it seems that visual experiences typically have detailed representational content. The same goes for experiences in other sensory modalities, and arguably for many or most nonsensory experiences as well.

Much of the most interesting work on NCCs is concerned with states like these. This is work on the neural correlates of the contents of consciousness. Much work on the neural correlates of visual consciousness has this character, for example. This work is not concerned merely with the neural states that determine that one *has* visual consciousness; it is concerned with the neural states that determine the specific contents of visual consciousness.

A nice example is supplied by the work of Logothetis and colleagues on the NCC of visual consciousness in monkeys (Logothetis and Schall 1989; Leopold and Logothetis 1996; Sheinberg and Logothetis 1997). In this work, a monkey is trained to press various bars when they are confronted with various sorts of images: horizontal and vertical gratings, for example, or gratings drifting left and right, or faces and sunbursts (I will use horizontal and vertical gratings for the purposes of illustration). After training is complete, the monkey is presented with two stimuli at once, one to each eye. In humans, this usually

produces binocular rivalry, with alternating periods of experiencing a definite image, and occasional partial overlap. The monkey responds by pressing bars, in effect "telling" the experimenter what it is seeing: a horizontal grating, or a vertical grating, or an interlocking grid.

At the same time, neurons in the monkey's cortex are being monitored by electrodes. It is first established that certain neurons respond to certain stimuli: to horizontal lines, for example, or to flowers. Then these neurons are monitored in the binocular rivalry situation, to see how well they correlate with what the monkey seems to be seeing. It turns out that cells in primary visual cortex (V1) don't correlate well: when the monkey is stimulated with horizontal and vertical gratings but "sees" horizontal, a large number of "vertical" cells in V1 fire, as well as "horizontal" cells. At this point, most cells seem to correlate with retinal stimulus, not with visual percept. But further into the visual system, the correlation increases, until in inferior temporal cortex, there is a very strong correlation. When the monkey is stimulated with horizontal and vertical grating but "sees" horizontal, almost all of the relevant horizontal cells in IT fire, and almost none of the vertical cells do. When the monkey's response switches, indicating that it is now "seeing" vertical, the cell response switches accordingly.

These results lend themselves naturally to speculation about the location of a visual NCC. It seems that V1 is unlikely to be or involve an NCC, for example, due to the failure of V1 cells to correlate with the contents of consciousness. Of course there are still the possibilities that some small subset of V1 is an NCC, or that V1 is a neural correlate of some aspects of visual consciousness but not of others, but I leave those aside for now. On the other hand, IT seems to be a natural candidate for the location of an NCC, due to the strong correlation of its cells with the content of consciousness. At least it is natural to suppose that IT is a "lower bound" on the location of a visual NCC (due to the failure of strong correlation before then), though the NCC itself may be further in. None of this evidence is conclusive (and Logothetis and colleagues are appropriately cautious), but it is at least suggestive.

It is clear that this work is concerned with the neural correlates of the *contents* of visual consciousness. We are interested in finding cortical areas whose neural activity correlates with and predicts specific contents of consciousness, such as experiences of horizontal or vertical lines, or of flowers or sunbursts. The ideal is to find a neural system from whose activity we might determine the precise contents of a visual experience, or at least of its contents in certain respects (shape, color, and the like).

Interestingly, it seems that in doing this we are crucially concerned with the the representational contents of the neural systems themselves. In the Logothetis work, for example, it is important to determine the receptive fields of the cells (whether they respond to horizontal or vertical gratings, for example), in order to see whether the receptive fields of active cells matches up with the apparent contents of visual consciousness. In essence, the receptive field is acting at least as a heuristic way of getting at representational content in the neurons in question. Then, the crucial question is whether the representational content in the neural system matches up with the representational content in visual consciousness.

This suggests a natural definition of a neural correlate of the contents of consciousness.

A neural correlate of the contents of consciousness is a neural representational system N such that representation of a content in N directly correlates with representation of that content in consciousness.

Or, more briefly:

A content NCC is a neural representational system N such that the content of N directly correlates with the content of consciousness.

For example, the Logothetis work lends itself to the speculation that IT might contain a content NCC for visual consciousness, since the content of cells in IT seems to directly correlate (at least in these experiments) with the contents of visual consciousness. (Much more investigation is required to see whether this correlation holds across the board, of course.)

This definition requires that we have some way of defining the representational content of a neural system independent of the contents of consciousness. There are various ways to do this. Using a cell's receptive field to define its representational content is probably the simplest. A more refined definition might also give a role to a system's projective field, and the sort of behavior that activity in that system typically leads to. And there may be more complex notions of representational content still, based on complex correlations with environment, patterns of behavior, and activity in other cells. But even a crude definition of representational content (e.g., the receptive field definition) is good enough for many purposes, and can yield informative results about the visual NCC.

It's arguable that much work on the visual NCC tacitly invokes this sort of definition. Another example is Milner and Goodale's work on the two pathways of visual perception. They suggest that the ventral stream is largely for cognitive identification and decision, while the dorsal stream is largely for online motor response, and the visual consciousness correlates with activity in the ventral stream.

Much of the support for this work lies with patients who have dissociations between specific contents of conscious perception and the contents involved in motor response. For example, a subject with visual form agnosia (e.g. Milner and Goodale's patient D.F.) cannot consciously identify a vertical slot, but can "post" an envelope through it without problem; while subjects with optic ataxia (e.g. those with Balint's (1909) syndrome) can identify an object but cannot act appropriately toward it. The dissociations here appear to go along with damage to the ventral and dorsal pathways respectively.

What seems to be going on, on a natural interpretation of these results and of Milner and Goodale's hypothesis, is that for these subjects, there is a dissociation between the contents represented in the ventral pathways and those represented in the dorsal pathway. In these cases, the character of a motor response appears to be determined by the contents represented in the dorsal pathway, but the character of conscious perception appears to be determined by the contents represented in the ventral pathway.

Thus one can see Milner and Goodale's hypothesis as involving the suggestion that the ventral stream

contains the neural correlates of the contents of visual consciousness. The hypothesis is quite speculative, of course (though it is interesting to note that IT lies in the ventral stream), but it seems that the content-based analysis provides a natural interpretation of what the hypothesis is implicitly claiming regarding the visual NCC, and of what may follow if the hypothesis turns out to be correct.

One could give a similar analysis of much or most work on the visual NCC. When Crick and Koch (1998) propose that the visual NCC lies outside V1, for example, much of the experimental evidence they appeal to involves cases where some content is represented in consciousness but not in V1, or vice versa. For example, Gur and Snodderly (1997) show that for some quickly alternating isoluminant color stimuli, color cells in V1 flicker back and forth even though a single fused color is consciously perceived. And results by He et al (1996) suggest that orientation of a grating can fade from consciousness even though orientation cells in V1 carry the information. The results are not entirely conclusive, but they suggest a mismatch between the representational content in V1 and the content of consciousness.

One can apply this sort of analysis equally to NCCs in other sensory modalities. An NCC of auditory consciousness, for example, might be defined as a neural representational system whose contents correlate directly with the contents of auditory consciousness: loudness, direction, pitch, tone, and the like. The idea can arguably apply to defining the neural correlates of bodily sensations, of conscious mental imagery, and perhaps of conscious emotion and of the stream of conscious thought. All these aspects of consciousness can be naturally analyzed (at least in part) in terms of their content. In looking for their respective NCCs, we may ultimately be looking for neural systems whose content correlates with the contents of these aspects of consciousness.

(iv) *Arbitrary phenomenal properties*

(This section is a little more technical than those above, and might be skipped by those not interested in philosophical details.)

One might try to give a general definition of an NCC of various states of consciousness, of which each of the above would be a special case. To do this, one would need a general way of thinking about arbitrary states of consciousness. Perhaps the best way is to think in terms of arbitrary *phenomenal properties*. For any distinctive kind of conscious experience, there will be a corresponding phenomenal property: in essence the property of having a conscious experience of that kind. For example, being in a hypnotic state of consciousness is a phenomenal property; having a visual experience of a horizontal line is a phenomenal property; feeling intense happiness is a phenomenal property; feeling a throbbing pain is a phenomenal property; being conscious is a phenomenal property. Phenomenal properties can be as coarse-grained or as fine-grained as you like, as long as they are wholly determined by the current conscious state of the subject.

With this notion in hand, one might try to define the neural correlate of an arbitrary phenomenal property P.

A state N1 of system N is a neural correlate of phenomenal property P if N's being in N1

directly correlates with the subject having P.

Note that we here talk of a *state* being an NCC. Given a *specific* phenomenal property - experiencing a horizontal line, for example, it is no longer clear that it makes sense to speak of a given system being the NCC of that property. Rather, it will be a particular state of that system. Neural firing in certain horizontal cells in IT (say) might be a neural correlate of seeing a horizontal line, for example; and having one's neurochemical system in a certain region of state space might be a neural correlate of waking consciousness, on Hobson's hypothesis. These are specific states of the neural systems in question.

Most of the time, we are not concerned with neural correlates of single phenomenal properties, but of *families* of phenomenal properties. Hobson is concerned not just with the neural correlate of waking consciousness, for example, but with the neural correlate of the whole family of background states of consciousness. Work on the visual NCC is not concerned with just the neural correlate of horizontal experience, but with the neural correlates of the whole system of visual experiential contents.

We might say a *phenomenal family* is a set of mutually exclusive phenomenal properties that jointly partition the space of conscious experiences, or at least some subset of that space. That is, any subject having an experience (of a certain relevant kind) will have a phenomenal property in the family, and will not have more than one such property. Specific contents of visual consciousness make for a phenomenal family, for example: any visually conscious subject will have some specific visual content, and they will not have two contents at once (given that we are talking about *overall* visual content). The same goes for contents at a particular location in the visual field: anyone with an experience as of a certain location will have some specific content associated with that location (a red horizontal line, say), and not more than one. (Ambiguous experiences are not counterexamples here, as long as we include ambiguous contents as members of the family in question.) The same again goes for color experience at any given location: there will be a phenomenal family (one property for each color quality) for any such location. And the same goes for background states of consciousness. All these sets of phenomenal properties make phenomenal families.

We can then say:

A neural correlate of a phenomenal family S is a neural system N such that the state of N directly correlates with the subject's phenomenal property in S.

For any phenomenal family S, a subject will have at most one property in S (one background state, or one overall state of visual consciousness, or one color quality at a location). Neural system N will be an NCC of S when there are a corresponding number of states of N, one for every property in P, such that N's being in a given state directly correlates with the subject's having the corresponding phenomenal property. This template can be seen to apply to most of the definitions given above.

For the neural correlate of creature consciousness, we have a simple phenomenal family with two properties: being conscious and not being conscious. An NCC here will be a system with two states that

correlate with these two properties.

For the neural correlate of a background state of consciousness, we have a phenomenal family with a few more properties: dreaming, being in an ordinary waking state, being under hypnosis, etc. An NCC here will be a neural system with a few states that correlate directly with these properties. Hobson's neurochemical system would be an example.

For the neural correlate of contents of consciousness, one will have a much more complex phenomenal family (overall states of visual consciousness, or states of color consciousness at a location, or particular conscious occurrent thoughts), and a neural representational system to match. The state of the NCC will directly correlate with the specific phenomenal property.

Notice that in the content case, there is an extra strong requirement on the NCC. In the other cases, we have accepted an arbitrary match of neural states to phenomenal states - any state can serve as the neural correlate of a dreaming state of background consciousness, for example. But where content is concerned, not any neural state will do. We require that the *content* of the neural state in question must match the content of consciousness. This is a much stronger requirement.

It is arguable that this requirement delivers much greater explanatory and predictive power in the case of neural correlates of conscious content. The systematicity in the correlation means that it can be extended to predict the presence or absence of phenomenal features that may not have been present in the initial empirical data set, for example. And it also will dovetail more nicely with finding a mechanism and a functional role for the NCC that matches the role that we associate with a given conscious state.

It is this systematicity in the correlation that makes the current work on neural correlate of visual consciousness particularly interesting. Without it, things would be much more untidy. Imagine that we find arbitrary neural states that correlated directly with the experience of horizontal lines, etc, such that there was no corresponding representational content in the neural state. Instead, we match seemingly arbitrary states N1 with horizontal, N2 with vertical, and so on. Would we count this as a neural correlate of the contents of visual consciousness? If we did, it would be in a much weaker sense, and in a way that would lead to much less explanatory and predictive power.

One might then hope to extend this sort of systematicity to other, non-content-involving phenomenal families. For example, one might find among background states of consciousness some pattern, or some dimension along which they systematically vary (some sort of intensity dimension, for example, or a measure of alertness). If we could then find a neural system whose states do not just arbitrarily correlate with the phenomenal states in question, but which vary along a corresponding systematic dimension, then the NCC in question will have much greater potential explanatory and predictive power. So this sort of systematicity in phenomenal families is something that we should look for, and something that we should look to match in potential neural correlates.

Perhaps one could define a "systematic NCC" as a neural correlate of a phenomenal family such that states correlate with each other in some such systematic way. I will not try to give a general abstract

definition here, as things are getting complex enough already, but I think one can see a glimmer of how it might go. I will, however, keep using the case of neural correlate of the contents of consciousness (especially visual consciousness) as the paradigmatic example of an NCC, precisely because its definition builds in this such a notion of systematicity, with the corresponding explanatory and predictive power.

3 Direct correlation

The other thing that we need to clarify is the notion of "direct correlation". We have said that an NCC is a system whose state directly correlates with a state of consciousness, but what does direct correlation involve, exactly? Is it required that the neural system be necessary and sufficient for consciousness, for example, or merely sufficient? And over what range of cases must the correlation obtain for the system to count as an NCC? Any possible case? A relevantly constrained set of cases? And so on.

The paradigmatic case will involve a neural system N with states that correlate with states of consciousness. So we can say that

state of N -?- state of consciousness

and specifically

N is in state N1 -?- subject has conscious state C.

In the case of the contents of consciousness, we have a system N such that representing a content in N directly correlates with representation in consciousness. So we can say

representing C in N -?- representing C in consciousness.

The question in all these cases concerns the nature of the required relation. How strong a relation is required here for N to be a NCC?

(A) *Necessity, sufficiency?*

The first question is whether the NCC state is required to be necessary and sufficient for the conscious state, merely sufficient, or something else in the vicinity.

(A1) *Necessity and sufficiency.* The first possibility is that the state of N is necessary and sufficient for the corresponding state of consciousness. This is an attractive requirement for an NCC, but it is arguably too strong. It might turn out that there is more than one neural correlate of a given conscious state. For example, it may be there there are two systems, M and N, such that a certain state of M suffices for being in pain and a certain state of N also suffices for being in pain, where these two states are not themselves always correlated. In this case, it seems that we would likely say that both M and N (or their

corresponding states) are neural correlates of pain. But it is not the case that activity in M is necessary and sufficient for pain (as it is not necessary), and the same goes for N. If both M and N are to count as NCCs here, we cannot require an NCC to be necessary and sufficient.

(A2) *Sufficiency*. From the above, it seems plausible that we require only that an NCC state be *sufficient* for the corresponding state of consciousness, not necessary. But is any sufficient state enough? If it is, then it seems that the whole brain will count as an NCC of any state of consciousness. The whole brain will count as an NCC of pain, for example, since being in a certain total state of the whole brain will suffice for being in pain. Perhaps there is some very weak sense in which this makes sense, but it does not seem to capture what researchers in the field are after when looking for an NCC. So something more than mere sufficiency is required.

(A3) *Minimal sufficiency*. The trouble with requiring mere sufficiency, intuitively, is that it allows irrelevant processes into a NCC. If N is an NCC, then the system obtained by conjoining N with a neighboring system M will also qualify as an NCC by the previous definition, since the state of N+M will suffice for the relevant states of consciousness.

The obvious remedy is to require that an NCC has to be a *minimal sufficient system*: that is, a *minimal* system whose state is sufficient for the corresponding conscious state. By this definition, N will be an NCC when (1) the states of N suffice for the corresponding states of consciousness, and (2) no proper part M of N is such that the states of M suffice for the corresponding states of consciousness. In this way, we pare down any potential NCC to its core: any irrelevant material will be whittled away, and an NCC will be required to contain only the core processes that suffice for the conscious state in question.

Note that on this definition, there may be more than one NCC for a given conscious state. It may be that there is more than one minimal sufficient system for a given state, and both of these will count as a neural correlate of that state. The same goes for systems of phenomenal states. This seems to be the right result: we cannot know a priori that there will be only one NCC for a given state or system of states. Whether there will actually be one or more than one for any given state, however, is something that can be determined only empirically.

There is a technical problem for the minimality requirement. It may turn out that there is significant redundancy in a neural correlate of consciousness, such that for example a given conscious visual content is represented redundantly in many cells in a given area. If this is so, then that visual area as a whole might not qualify as a minimal sufficient system, as various smaller components of it might all themselves correlate with the conscious state. In this case the definition above would imply that various such small components would each be an NCC. One could deal with this sort of case by noting that the problem arises only when the states of the various smaller systems are themselves wholly correlated with each other. (If their mutual correlation can be broken, so will their correlation with consciousness, so that the overall system or some key subsystem will again emerge as the true NCC). Given this, one could stipulate that where states of minimal sufficient systems are wholly correlated with each other, it is the union of the system that should be regarded as an NCC, rather than the individual systems. So an NCC would be a minimal system whose state is sufficient for a given conscious state and whose state is not wholly

correlated with the state of any other system. I will pass over this complication in what follows, however.

(B) *What range of cases?*

An NCC will be a minimal neural system N such that the state of N is sufficient for a corresponding conscious state C. This is to say: if the system is in state N1, the subject will have conscious state C. But the question now arises: over what range of cases must the correlation in question hold?

There is sometimes a temptation to say that this question does not need to be answered: all that is required is to say that *in this very case*, neural state N1 suffices for or correlates with conscious state C. But this does not really make sense. There is no such thing as a single-case correlation. Correlation is always defined with respect to a range of cases. The same goes for sufficiency. To say that neural state N1 suffices for conscious state C is to say that in a range of cases, neural state N1 will always be accompanied by conscious state C. But what is the range of cases?

(B1) *Any possible case.* It is momentarily tempting to suggest that the correlation should range across any possible case: if N is an NCC it should be impossible to be in a relevant state of N without being in the corresponding state of consciousness. But a moment's reflection suggests that this is incompatible with the common usage in the field. NCCs are often supposed to be relatively limited systems, such as the inferior temporal cortex or the intralaminar nucleus. But nobody (or almost nobody) holds that if one excises the entire inferior temporal cortex or intralaminar nucleus and puts it in a jar, and puts the system into a relevant state, it will be accompanied by the corresponding state of consciousness.

That is to say, for a given NCC, it certainly seems *possible* that one can have the NCC state without the corresponding conscious state, for example by performing sufficiently radical lesions. So we cannot require that the correlation range over all possible cases.

Of course, one could always insist that a *true* NCC must be such that it is impossible to have the NCC state without the corresponding conscious state. The consequence of this would be that an NCC would almost certainly be far larger than it is on any current hypothesis, as we would have to build in a large amount of the brain to make sure that all the background conditions are in place. Perhaps it would be some sort of wide-ranging although skeletal brain state, involving aspects of processes from a number of regions of the brain. This might be a valid usage, but it is clear that this is not what researchers in the field are getting at when they are talking about an NCC.

We might call the notion just defined a *total* NCC, as it builds in the totality of physical processes that are absolutely required for a given conscious state. The notion that is current in the field is more akin to that of a *core* NCC. (I adapt this terminology from Shoemaker's (1979) notion of a "total realization" and a "core realization" of a functional mental state.) A total NCC builds in everything and thus automatically suffices for the corresponding conscious states. A core NCC, on the other hand, contains only the "core" processes that correlate with consciousness. The rest of the total NCC will be relegated to some sort of background conditions, required for the correct functioning of the core.

(Philosophical note: The sort of possibility being considered here is natural or nomological possibility, or possibility compatible with the laws of nature. If we required correlation across all *logically* possible cases, there might be no total NCC at all, as it is arguably logically possible (or coherently conceivable) to instantiate any physical process at all without consciousness. But it is probably not naturally possible. It is almost certainly naturally necessary that a being with my brain state will have the same sort of conscious state as me, for example. So natural possibility and necessity is the relevant sort for defining the correlation here.)

The question is then how to distinguish the core from the background. It seems that what is required for an NCC (in the "core" sense) is not that it correlate with consciousness across any possible conditions, but rather that it correlate across some constrained range of cases in which some aspects of normal brain functioning are held constant. The question then becomes, what is to be held constant? Across just what constrained range of cases do we require that an NCC correlate with consciousness?

(B2) *Ordinary functioning brain in ordinary environments.*

One might take the moral of the above to be that one cannot require an NCC to correlate with consciousness in "unnatural" cases. What matters is that the NCC correlates with consciousness in "natural" cases, those that actually occur in the functioning of a normal brain. The most conservative strategy would be to require correlation only across cases involving a normally functioning brain in a normal environment, receiving "ecologically valid" inputs of the sort received in a normal life.

The trouble with this criterion is that it seems too weak to narrow down the NCC. It may turn out that this way, we find NCCs at all stages of the visual system, for example. In normal visual environment, we can expect that the contents of visual systems from V1 through IT will all correlate with the contents of visual consciousness, and that even the contents of the retina will to some extent. The reason is that in normal cases all these will be linked in a straightforward causal chain, and the systems in question will not be dissociated. But it seems wrong to say that merely because of this, all the systems (perhaps even the retina) should count as an NCC.

The moral of this is that we need a more fine-grained criterion to dissociate these systems and to distinguish the core NCC from processes that are merely causally linked to it. To do this, we have to require correlation across a range of *unusual cases* as well as across normal cases, as it is these cases that yield interesting dissociations.

(B3) *Normal brain, unusual inputs.* The next most conservative suggestion is that we still require a normal brain for our range of cases, but that we allow any possible inputs, including "ecologically invalid" inputs. This would cover the Logothetis experiments, for example. The inputs that evoke binocular rivalry are certainly unusual, and not encountered in a normal environment. But it is precisely these that allow the experiments to make more fine-grained distinctions than we normally can. The experiments suggest that IT is more likely than V1 to be an NCC precisely because it correlates with consciousness across the wider range of cases. If states of V1 truly do not match up with states of

consciousness in this situation, then it seems that V1 cannot be an NCC. If that reasoning is correct, then it seems that we require an NCC to correlate with consciousness across all unusual inputs, and not just across normal environments.

The extension of the correlation requirement from normal environments to unusual inputs is a relatively "safe" extension and seems a reasonable requirement, though those who place a high premium on ecological validity might contest it. But it is arguable that this is still too weak to do the fine-grained work in distinguishing an NCC from systems linked to it. Presumably unusual inputs will go only so far in yielding interesting dissociations, and some systems (particularly those well down the processing pathway) may well stay associated on any unusual inputs. So it is arguable that we will need more fine-grained tools to distinguish the NCC.

(B4) *Normal brain, vary brain stimulation.* The next possibility is to allow cases involving not just unusual inputs, but involving direct stimulation of the brain. Such direct stimulation might include both electrode stimulation and transcranial magnetic stimulation. On this view, we will require that an NCC correlates with consciousness across all cases of brain stimulation, as well as normal functioning. So if we have a potential NCC state that does not correlate with consciousness in the right way in a brain stimulation condition, that state will not be a true NCC.

This requirement seems to fit some methods used in the field. Penfield (e.g. Penfield and Rasmussen 1950) pioneered the use of brain stimulation to draw conclusions about the neural bases of consciousness. Libet (1982) has also used brain stimulation to good effect, and more recently Newsome and colleagues (e.g. Salzman et al 1990) have used brain stimulation to draw some conclusions about neural correlates of motion perception in monkeys. (See also Marge 1991 for a review of transcranial magnetic stimulation in vision.)

Brain stimulation can clearly be used to produce dissociations more fine-grained than can be produced merely with unusual inputs. One might be able to dissociate activity in any system from that in a preceding system by stimulating that system directly, for example, as long as there are not too many backwards connections. Given a candidate NCC - inferior temporal cortex, say - one can test the hypothesis by stimulating an area immediately following the candidate in the processing pathway. If that yields a relevant conscious state without relevant activity in IT (say), that indicates that IT is probably not a true NCC after all. Rather, the NCC may lie in a system further down the processing chain. (I leave aside the possibility that there might be two NCCs at different stages of the chain.)

This reasoning seems sound, suggesting that we may tacitly require an NCC to correlate with consciousness across brain stimulation conditions. There is no immediately obvious problem with the requirement, at least when the stimulation in question is relatively small and localized. If one allows arbitrary large stimulation, there may be problems. For example, one could presumably use brain stimulation at least in principle to disable large areas of the brain (by overstimulating those areas, for example) while leaving NCC activity intact. In this case, it is not implausible to expect that one will have the relevant NCC activity without the usual conscious state (just as in the case where one lesions the whole NCC and puts it in a jar), so the correlation will fail in this case. But intuitively, this does not seem

to disprove the claim that the NCC in question is a true NCC, at least before the stimulation. If so, then we cannot allow unlimited brain stimulation in the range of cases relevant to the correlation; and more generally, some of the problems for lesions (discussed below) may apply to reasoning involving brain stimulation. Nevertheless, one might well require that an NCC correlate with consciousness at least across cases of limited stimulation, in the absence of strong reason to believe otherwise.

(B5) *Abnormal functioning, due to lesions.* In almost all of the cases above, we have retained a normally functioning brain; we have just stimulated it in unusual ways. The next logical step is to allow cases where the brain is not functioning normally, due to lesions in brain systems. Such lesions might be either natural (e.g. due to some sort of brain damage) or artificial (e.g. induced by surgery). On the latest view, we will require that an NCC correlates with states of consciousness not just over cases of normal functioning, but over cases of abnormal functioning as well.

This certainly squares with common practice in the field. Lesion studies are often used to draw conclusions about the neural correlates of consciousness. In Milner and Goodale's work, for example, the fact that consciousness remains much the same upon lesions to the dorsal stream but not to the ventral stream is used to support the conclusion that the NCC lies within the ventral stream. More generally, it is often assumed that if some aspect of consciousness survives relatively intact with a given brain area is damaged, then that brain area is unlikely to be or contain an NCC.

The tacit premise in this research is that an NCC should correlate with consciousness but just in cases of normal functioning but in cases of abnormal functioning as well. Given this premise, it follows that if we find an abnormal case in which neural system N is damaged but a previously corresponding conscious state C is preserved, then N is not a neural correlate of C. Without this premise, or a version of it, it is not clear that any such conclusion can be drawn from lesion studies.

The premise may sound reasonable, but we already have reason to be suspicious of it. We know that for any candidate NCC, sufficiently radical changes can destroy the correlation. Preserving merely system N, cut off from the rest of the brain, for example, is unlikely to yield a corresponding conscious state; but intuitively, this does not imply that N was not an NCC in the original case.

Less radically, one can imagine placing lesions immediately downstream from a candidate NCC N, so that N's effects on the rest of the brain are significantly reduced. In such a case, it is probable that N can be active without the usual behavioral effects associated with consciousness, and quite plausibly without consciousness itself. It's not implausible that an NCC supports consciousness largely in virtue of playing the right functional role in the brain; by virtue of mediating global availability, for example (see Baars 1988 and Chalmers 1998). If so, then if the system is changed so that the NCC no longer plays that functional role, then NCC activity will no longer correlate with consciousness. But the mere fact that correlation can be destroyed by this sort of lesion does not obviously imply that N is not an NCC in a normal brain. If that inference could be made, then almost any candidate NCC could be ruled out by the right sort of lesion.

It may be that even smaller lesions can destroy a correlation in this way. For example, it is not implausible

that for any candidate NCC N, there is some other local system in the brain (perhaps a downstream area) whose proper functioning is required for activity in the N to yield the usual effects that go with consciousness, and for N to yield consciousness itself. This second system might not itself be an NCC in any intuitive sense; it might merely play an enabling role, in the way that proper functioning of the heart plays an enabling role for functioning of the brain. If so, then if one lesions this single area downstream, then activity in the N will no longer correlate with consciousness. In this way, any potential NCC might be ruled out by a localized lesion elsewhere.

The trouble is that lesions change the architecture of the brain, and it's quite possible that changes to brain architecture can change the very location of an NCC, so that a physical state that was an NCC in a normal brain will not be an NCC in the altered brain. Given this possibility, it seems too strong to require that an NCC correlate with consciousness across arbitrary lesions and changes in brain functioning. We should expect an NCC to be architecture-dependent, not architecture-independent.

So an NCC should not be expected to correlate with consciousness across arbitrary lesion cases. There are now two alternatives. Either we can require correlation across some more restricted range of lesion cases, or one can drop the requirement of correlation in abnormal cases altogether.

For the first alternative to work, we would have to find some way to distinguish a class of "good" lesions from the class of "bad" lesions. An NCC would be expected to correlate with consciousness across the good lesions but not the bad lesions. If one found a "good" lesion case where activity in system N was present without the corresponding consciousness state, this would imply that N is not an NCC; but no such conclusion could be drawn from a "bad" lesion case.

The trouble is that it is not at all obvious that such a distinction can be drawn. It might be tempting to come up with an after-the-fact distinction, defined as the range of lesions in which correlation with a given NCC N is preserved, but this will not be helpful at all, as we are interested in precisely the criterion that makes N qualify as an NCC in the first place. So a distinction will have to be drawn on relatively a priori grounds (it can then be used to determine whether a given correlation-pattern qualifies an arbitrary system as an NCC or not). But it is not clear how to draw the distinction. One might suggest that correlation should be preserved across small lesions but not large ones; but we have seen above that even small lesions might destroy a potential NCC. Or one might suggest that lesions in downstream areas are illegitimate, but upstream and parallel lesions are legitimate. But even here, it is not clear that indirect interaction with an upstream or parallel area might be required to support proper functioning of an NCC. Perhaps with some ingenuity one might be able to come up with a criterion, but it is not at all obvious how.

The second alternative is to hold that correlation across cases of normal functioning (perhaps with unusual inputs and brain stimulation) is all that is required to be an NCC. If this is so, one can never infer directly from the fact that N fails to correlate with consciousness in a lesion case to the conclusion that N is not an NCC. On this view, the location of an NCC is wholly architecture-dependent, or entirely dependent on the normal functioning of the brain. One cannot expect an NCC to correlate with consciousness in cases with abnormal functioning or different architecture, so no direct conclusion can be drawn from failure of

correlation across lesion cases. Of course, one can still appeal to cases with unusual inputs and brain stimulation to make fine-grained distinctions among NCCs.

The main conceptual objection to the second alternative is that one might *need* lesion cases to make the most fine-grained distinctions that are required. Consider a hypothetical case in which we have two linked systems N and M which correlate equally well with consciousness across all normal cases, including all unusual inputs and brain stimulation, but such that in almost all relevant lesion cases, consciousness correlates much better with N than with M. In this case, might we want to say that N rather than M is an NCC? If so, we have to build in some allowance for abnormal cases into the definition of an NCC. An advocate of the second alternative might reply by saying that such cases will be very unusual, and that if N and M are dissociable by lesions, there is likely to be some unusual brain stimulation that will bring out the dissociation as well. In the extreme case where no brain stimulation leads to dissociation, one might simply bite the bullet and say that both N and M are equally good NCCs.

Taking everything into consideration, I am inclined to think the second alternative is better than the first. It seems right to say that "core" NCC location depends on brain architecture and normal functioning, and it is unclear that correlation across abnormal cases should be required, especially given all the associated problems. A problem like the one just mentioned might provide some pressure to investigate the first alternative further, and I do not rule out the possibility that some way of distinguishing "good" from "bad" lesions might be found, but all in all it seems best to say that an NCC cannot be expected to correlate with consciousness across abnormal cases.

Of course this has an impact on the methodology in the search for an NCC. As we have seen, lesion studies are often used to draw conclusions about NCC location (as in the Milner and Goodale research, for example, and also in much research on blindsight), and failure of correlation in lesion cases is often taken to imply that a given system is not an NCC. But we have seen that the tacit premise of this sort of research - that an NCC must correlate across abnormal as well as normal cases - is difficult to support, and leads to significant problems. So it seems that lesion studies are methodologically dangerous here. One should be very cautious in using them to draw conclusions about NCC location.

This is not to say that lesion studies are irrelevant in the search for an NCC. Even if correlation across abnormal cases is not *required* for system N to be an NCC, it may be that correlation across abnormal cases can provide good *evidence* that N is an NCC, and that failure of such correlation in some cases provides good evidence that N is not an NCC. Say we take the second alternative above, and define an NCC as a system that correlates with consciousness across all normal cases (including unusual input and stimulation). It may nevertheless be the case that information about correlations across all these normal cases with unusual stimulation is difficult to come by (due to problems in monitoring brain systems at a fine grain, for example), and that information about correlation across lesion cases is easier to obtain. In this case, one might sometimes take correlation across abnormal cases as *evidence* that a system will correlate across the normal cases in question, and thus as evidence that the system is an NCC. Similarly, one might take failure of correlation across abnormal cases as evidence that a system will fail to correlate across certain normal cases, and thus as evidence that the system is not an NCC.

The question of whether a given lesion study can serve as evidence in this way needs to be taken on a case-by-case basis. It is clear that some lesion studies will not provide this sort of evidence, as witnessed by the cases of severe lesions and downstream lesions discussed earlier. In the cases, failure of correlation across abnormal cases provides no evidence of failure of correlation across normal cases. On the other hand, it does not seem unreasonable that the Milner and Goodale studies should be taken as evidence that even in normal cases, the ventral stream will correlate better with visual consciousness than the dorsal stream. Of course the real "proof" would come from a careful investigation of the relevant processes across a wide range of "normal" cases involving both standard environments, unusual inputs, and brain stimulation; but in the absence of such a demonstration, the lesion cases at least provide suggestive evidence.

In any case, the moral is that one has to be very cautious when drawing conclusions about NCC location from lesion studies. At best these studies serve as indirect evidence rather than as direct criteria, and even as such there is a chance that the evidence can be misleading. One needs to consider the possibility that the lesion in question is changing brain architecture in such a fashion that what was once an NCC is no longer an NCC, and one needs to look very closely at what is going on to rule out the possibility. It may be that this can sometimes be done, but it is a nontrivial matter.

4 Overall definition

With all this, we have come to a more detailed definition of an NCC. The general case is something like the following:

An NCC is a minimal neural system N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient, under conditions C , for the corresponding state of consciousness.

The central case of the neural correlate of the content of consciousness can be put in more specific terms.

An NCC (for content) is a minimal neural representational system N such that representation of a content in N is sufficient, under conditions C , for representation of that content in consciousness.

One might also give a general definition of the NCC for an arbitrary phenomenal property or for a phenomenal family, but I will leave those aside here.

The "conditions C " clause here represents the relevant range of cases, as discussed above. If the reasoning above is on the right track, then conditions C might be seen as conditions involving normal brain functioning, allowing unusual inputs and limited brain stimulation, but not lesions or other changes in architecture. Of course the precise nature of conditions C is still debatable. Perhaps one could make a case for including a limited range of lesion cases in the definition. In the other direction, perhaps one might make a case that the requirement of correlation across brain stimulation or unusual inputs is too strong, due to the abnormality of those scenarios. But I think the conditions C proposed here are at least a

reasonable first pass, pending further investigation.

Of course, to some extent, defining what "really" counts as an NCC is a terminological matter. One could quite reasonably say that there are multiple different notions of NCC, depending on just how one understands the relevant conditions C, or the matter of necessity and sufficiency, and so on; and not much really rests on which of these is the "right" definition. Still, we have seen that different definitions give very different results, and that many potential definitions have the consequence that systems that intuitively seem to qualify as an NCC do not qualify after all, and that NCC hypotheses put forward by researchers in the field could be ruled out on trivial a priori grounds. Those consequences seem undesirable. It makes sense to have a definition of NCC that fits the way the notion is generally used in the field, and that can make sense of empirical research in the area. At the same time we want a definition of NCC to be coherent and well-motivated in its own right, such that an NCC is something worth looking for, and such that the definition can itself be used to assess various hypotheses about the identity of an NCC. It seems to me that the definition I have given here is at least a first pass in this direction.

5 Methodological consequences

The discussion so far has been somewhat abstract, and the definitions given above may look like mere words, but from these definitions and the reasoning that went into them, one can straightforwardly extract some concrete methodological recommendations for the NCC search. Many of these recommendations are plausible or obvious in their own right, but it is interesting to see them emerge from the analysis.

(i) *Lesion studies are methodologically dangerous.* Lesion studies are often used to draw conclusions about neural correlates of consciousness, but we have seen that their use can be problematic. The identity of an NCC is arguably always relative to specific brain architecture and normal brain functioning, and correlation across abnormal cases should not generally be expected. In some cases, lesion studies can change brain architecture so that a system that was previously an NCC is no longer an NCC. So one can never infer directly from failure of correlation between a system and consciousness in a lesion case to the conclusion that that system is an NCC. Sometimes one can infer this indirectly, by using the failure of correlation here as evidence for failure of correlation in normal cases, but one must be cautious.

(ii) *There may be many NCCs.* On the definition above, an NCC is a system whose activity is *sufficient* for certain states of consciousness. This allows for the possibility of multiple NCCs, in at least two ways. First, different sorts of conscious states may have different corresponding NCCs; there may be different NCC for visual and auditory consciousness, for example, and perhaps even for different aspects of visual consciousness. Second, even for a particular sort of conscious state (such as pain), we cannot rule out the possibility that there will be two different systems whose activity is sufficient to produce that state.

Of course it *could* turn out that there is only a small number of NCCs, or perhaps even one. For all that I have said here, it is possible that there is some central system which represents the contents of visual consciousness, auditory consciousness, emotional experience, the stream of conscious thought, the background state of consciousness, and so on. Such a system might be seen as a sort of "consciousness

module", or perhaps as a "Cartesian theater" (Dennett 1991) or a "global workspace" (Baars 1988), depending on whether one is a foe or a friend of the idea (see Chalmers 1998 for some discussion). But it is by no means obvious that there will be such a system, and I think the empirical evidence so far is against it. In any case, the matter cannot be decided a priori, so our definition should be compatible with the existence of multiple or many NCCs.

(3) *Minimize size of an NCC.* We have seen that an NCC should be understood as a *minimal* neural system that correlates with consciousness. Given this, we should constrain the search for the NCC by aiming to find a neural correlate that is as small as possible. Given a broad system that appears to correlate with consciousness, we need to isolate the core relevant parts and aspects of that system that underlie the correlation. And given the dual hypotheses that consciousness correlates with a broad system or a narrower system contained within it, we might first investigate the "narrow" hypothesis, as if it correlates with consciousness, the broad system cannot be a true NCC.

So to some extent it makes sense to "start small" in the search for an NCC. This fits the working methodology proposed by Crick and Koch (1998). Crick and Koch suggest that an NCC may perhaps involve a very small number of neurons (perhaps in the thousands) with certain distinctive properties. There is no guarantee that this is correct (and my own money is against it), but it makes a good working hypothesis in the NCC search. Of course one should simultaneously investigate broad systems for correlation with consciousness, so that one can then focus on those areas and try to narrow things down.

(4) *Distinguish NCC for background state and for content.* We have seen that there may be different NCCs for different sorts of states of consciousness. An important distinction in this class is that between the neural correlate of background state of consciousness (wakefulness, dreaming, etc) and the neural correlate of specific contents. It may be that these are quite different systems. It is not implausible on current evidence that an NCC for background state involves processes in the thalamus, or thalamocortical interactions, while an NCC for specific contents of consciousness involves processes in the cortex. These different sorts of NCC will require quite different methods for their investigation.

(5) *NCC studies need to monitor neural representational content.* Arguably the most interesting part of the NCC search is the search for neural determinants of specific contents of consciousness, such as the contents of visual consciousness. We have seen that an NCC here will be a neural representational system whose contents are correlated with the contents of consciousness. To determine whether such a system is truly an NCC, then, we need methods that monitor the representational content of the system. This is just what we find in Logothetis's work, for example, where it is crucial to keep track of activity in neurons with known receptive fields.

This gets at a striking aspect of the NCC search in practice, which is that the most informative and useful results usually come from single-cell studies on monkeys. Large claims are sometimes made for brain imaging on humans, but it is generally difficult to draw solid conclusions from such studies, especially where an NCC is concerned. We can trace the difference to the fact that single-cell studies can monitor representational content in neural systems, whereas imaging studies cannot (or at least usually do not). The power of single-cell studies in the work of Logothetis, Andersen, Newsome and colleagues (e.g. the

works of Logothetis and Newsome already cited, and Bradley, Chang, and Andersen 1998) comes precisely from the way that cells can be monitored to keep track of their activity profile of neurons with known representational properties, such as receptive and projective fields. This allows us to track representational content in these neural systems and to correlate it with the apparent contents of consciousness. This is much harder to do in a coarse-grained brain imaging study, which generally tells one that there is an activity in a region while saying nothing about specific contents.

A moral is that it makes sense to concentrate on developing methods that can track neural representational content, especially in humans (where invasive studies are much more problematic, but where evidence for conscious content are much more straightforward). There has been some recent work on the using imaging methods to get at certain aspects of the content of visual consciousness, such as colors and shapes in the visual field (e.g. Engel et al 1997), and different sorts of objects that activate different brain areas (e.g. Tong et al 1998). There is also some current work on using invasive methods in neurosurgery patients to monitor the activity of single cells. One can speculate that if a noninvasive method for monitoring single-cell activity in humans is ever developed, the search for an NCC (like most of neuroscience) will be transformed almost beyond recognition.

(6) *Correlation across a few situations is limited evidence.* According to the definition above, an NCC is a system that correlates with consciousness across arbitrary cases of normal functioning, in any environment, with any unusual input or limited brain stimulation. In practice, though, evidence is far weaker than this. Typically one has a few cases, involving either a few subjects with different lesions, or study in which subjects are given different stimuli, and one notes an apparent correlation. This is only to be expected, given the current technological and ethical constraints on experimental methods. But it does mean that the evidence that current methods give is quite weak. To truly demonstrate that a given system is an NCC, one would need to demonstrate correlation across a far wider range of cases than is currently feasible. Of course current methods may give good *negative* evidence about systems that fail to correlate and thus are not NCCs, but strong positive evidence is harder to find. Positive hypotheses based on current sorts of evidence should probably be considered suggestive but highly speculative.

(7) *We need good criteria for the ascription of consciousness.* To find an NCC, we need to find a neural system that correlates with certain conscious states. To do this, we first need a way to know when a system is in a given conscious state. This is famously problematic, given the privacy of consciousness and the philosophical problem of other minds. In general, we rely on indirect criteria for the ascription of consciousness. The most straightforward of these criteria is verbal report in humans, but other criteria are often required. Where nonhuman subjects are involved, one must rely on quite indirect behavioral signs (voluntary bar-pressing in Logothetis's monkeys, for example).

A deep problem for the field is that our ultimate criteria here are not experimentally testable, as the results of any experiment will itself requires such criteria for its interpretation. (First-person experimentation on oneself as subject may be an exception, but even this has limitations.) So any experimental work implicitly relies on pre-empirical principles (even "philosophical" principles) for its interpretation. Given this, it is vital to refine and justify these pre-empirical principles as well as we can. In the case of verbal report, we may be on relatively safe ground (though even here there may be some grounds for doubt, as

witnessed in the debates over "subjective threshold" criteria in unconscious perception research; see e.g. Merikle and Reingold (1992). In other cases, especially nonhuman cases, careful attention to the assumptions involved here are required. I don't think this problem is insurmountable, but it deserves careful attention. Our conclusions about NCC location will be no better than the pre-experimental assumptions that go into the search. (I consider this problem, and its consequences for the NCC search, in much more detail in Chalmers 1998.)

Methodological summary: We can use all this to sketch a general methodology for the NCC search. First, we need methods for determining the contents of conscious experience in a subject, presumably by indirect behavioral criteria or by first-person phenomenology. Second, we need methods to monitor neural states in a subject, and in particular to monitor neural representational contents. Then we need to perform experiments in a variety of situations to determine which neural systems correlate with conscious states and which do not. Experiments involving normal brain functioning with unusual inputs and limited brain stimulation are particularly crucial here. Direct conclusions cannot be drawn from systems with lesions, but such systems can sometimes serve as indirect evidence. We need to consider multiple hypotheses in order to narrow down a set of minimal neural systems that correlate with consciousness across all relevant scenarios. We may well find many different NCCs in different modalities, and different NCCs for background state and conscious contents, although it is not out of the question that there will be only a small number. If all goes well, we might expect to eventually isolate systems that correlate strongly with consciousness across any normally functioning brain.

6 Should we expect an NCC?

One might well ask: given the notion of an NCC as I have defined it, is it guaranteed that there will *be* a neural correlate of consciousness?

In answering, I will assume that states of consciousness depend systematically in some way on overall states of the brain. If this assumption is false, as is held by some Cartesian dualists (e.g. Eccles 1994) and some phenomenal externalists (e.g. Dretske 1995), then there may be no NCC as defined here, as any given neural state might be instantiated without consciousness. (Even on these positions, an NCC *could* be possible, if it were held that brain states at least correlate with conscious states in ordinary cases). But if the assumption is true, then there will at least be some minimal correlation of neural states with consciousness.

Does it follow that there will be an NCC as defined here? This depends on whether we are talking about neural correlates of arbitrary conscious states, or about the more constrained case of neural correlates of conscious contents. In the first case, it is guaranteed that the brain as a whole will be a neural system which has states that suffice for arbitrary conscious states. So the brain will be one system whose state is sufficient for a given conscious state; and given that there is at least one such system for a given state, there must be at least one such *minimal* system for that state. Such a system will be an NCC for that state. Of course this reasoning does not guarantee that there will be only one NCC for a given state, or that the NCC for one state will be the same as the NCC for another, but we know that an NCC will exist.

In the case of neural correlates of the content of consciousness, things are more constrained, since a neural correlate is required not just to map to a corresponding state of consciousness, but to match it in *content*. This rules out the whole brain as even a non-minimal neural correlate, for example, since representing a content in the brain does not suffice to represent that content in consciousness (much of the brain's representational content is unconscious). Of course we may hope that there will be more constrained neural systems whose content systematically matches the contents of some aspect of consciousness. But one might argue that it is not obvious that such a system *must* exist. It might be held, for example, that the contents of consciousness are an emergent product of the contents of various neural systems, which together suffice for conscious content in question, but none of which precisely mirrors the conscious content.

I think one can plausibly argue that there is reason to expect that conscious contents will be mirrored by the contents of a neural representational system at *some* level of abstraction. In creatures with language, for example, conscious contents correspond well with contents that are made directly available for verbal report; and in conscious creatures more generally, one can argue that the contents of consciousness correspond to contents that are made directly available for the global voluntary control of behavior (see e.g. Chalmers 1998). So there is a correlation between the contents of consciousness and contents revealed or exhibited in certain functional roles within the system.

Given that these contents are revealed in verbal report and are exhibited in the control of behavior, there is reason to believe that they are represented at some point within the cognitive system. Of course this depends to some extent on just what "representation" comes to. On some highly constrained notions of representation - if it is held that the only true representation is symbolic representation, for example - then it is far from clear that the content revealed in behavior must be represented. But on less demanding notions of representation - on which, for example, systems are assigned representational content according to their functional role - then it will be natural to expect that the content revealed in a functional role will be represented in a system that plays that functional role.

This does not guarantee that there will be any single neural system whose content always matches the content of consciousness. It may be that the functional role in question is played by multiple systems, and that a given system may sometimes play the role, and sometimes not. If this is so, we may have to move to a higher level of abstraction. If there is no localizable neural system that qualifies as a correlate of conscious content, we may have to look at a more global system - the "global availability" system, for example, whereby contents are made available for report and global control - and argue that the contents of consciousness correspond to the contents made available in this system. If so, it could turn out that what we are left with is more like a "cognitive correlate of consciousness" (CCC?), since the system may not correspond to any neurobiological system whose nature and boundaries are independently carved out. But it can still function as a correlate in some useful sense.

In this context, it is important to note that an NCC need not be a specific anatomical area in the brain. Some of the existing proposals regarding NCCs involve less localized neurobiological properties. For example, Libet (1994) argues that the neural correlate of consciousness is temporally extended neural firing; Crick and Koch (1998) speculate that the NCC might involve a particular sort of cell, throughout

the cortex; Edelman (1988) suggests that the NCC might involve re-entrant thalamocortical loops; and so on. In these cases, NCCs are individuated by temporal properties, or by physiological rather than anatomical properties, or by functional properties, among other possibilities. If so, the "neural representational system" involved in defining a neural correlate of conscious content might also be individuated more abstractly: the relevant neural representational contents might be those represented by temporally extended firings, or by certain sorts of cells, or by re-entrant loops, and so on. So abstractness and failure of localization is not in itself a bar to a system's qualifying as an NCC.

It seems, then, that there are a range of possibilities for the brain-based correlates of conscious states, ranging from specific anatomical areas, through more abstract neural systems, to purely "cognitive" correlates such as Baars' (1988) "global workspace". Just how specific an NCC may turn out to be is an empirical question. One might reasonably expect that there will be some biological specificity. Within a given organism or species, one often finds a close match between specific functions and specific physiological systems, and it does not seem unlikely that particular neural systems and properties in the brain should be directly implicated in the mechanisms of availability for global control. If so, then we may expect specific neural correlates even of conscious contents. If not, we may have to settle for more abstract correlates, individuated at least partly at the cognitive level, though even here one will expect that some neural systems will be much more heavily involved than others. In any case it seems reasonable to expect that we will find informative brain-based correlates of consciousness at some level of abstraction in cognitive neurobiology.

Some have argued that we should not expect neural correlates of consciousness. For example, in their discussion of neural "filling-in" in visual perception, Pessoa, Thompson, and Noe (1998) argue against the necessity of what Teller and Pugh (1983) call a "bridge locus" for perception, which closely resembles the notion of a neural correlate of consciousness. Much of their argument is based on the requirement that such a locus must involve a spatiotemporal isomorphism between neural states and conscious states (so a conscious representation of a checkerboard would require a neural state in a checkerboard layout, for example). These arguments do not affect neural correlates of conscious contents as I have defined them, since a match between neural and conscious content does not require such a spatiotemporal correspondence (a neural representation of a shape need not itself have that shape). Pessoa et al also argue more generally against a "uniformity of content" thesis, holding that one should not expect a match between the "personal" contents of consciousness and the "subpersonal" contents of neural systems. I agree that the existence of such a match is not automatic, but as above I think that the fact that conscious contents are mirrored in specific functional roles gives reason to believe that they will be subpersonally represented at least at some level of abstraction.

It has also been argued (e.g. by Güzelde 1999) that there is probably no neural correlate of consciousness, since there is probably no area of the brain that is specifically dedicated to consciousness as opposed to vision, memory, learning, and so on. One may well agree that there is no such area, but it does not follow that there is no neural correlate of consciousness as defined here. An NCC (as defined here) requires only that a system be correlated with consciousness, not that it be dedicated solely or mainly to consciousness. The alternative notion of an NCC is much more demanding than the notion at issue in most empirical work on the subject, where it is often accepted that an NCC may be closely bound

up with visual processing (e.g. Logothetis, Milner and Goodale), memory (e.g. Edelman), and other processes. This becomes particularly clear once one gives up on the requirement that there be a single NCC, and accepts that there may be multiple NCCs in multiple modalities.

7 Conclusion

The discussion in the previous section helps bring out what an NCC is *not*, or at least what it might turn out not to be. An NCC is defined to be a *correlate* of consciousness. From this, it does not automatically follow that an NCC will be a system solely or mainly dedicated to consciousness, or even that an NCC will be the brain system most responsible for the generation of consciousness. It certainly does not follow that an NCC will itself yield an explanation of consciousness, and it is not even guaranteed that identifying an NCC will be the key to understanding the processes underlying consciousness. If we were to define an NCC in these stronger terms, it would be far from obvious that there must be an NCC, and it would also be much less clear how to search for an NCC.

Defining an NCC in terms of correlation seems to capture standard usage best, but it also makes the search more clearly defined, and makes the methodology clearer. Correlations are easy for science to study. It also means that the search for an NCC can be to a large extent theoretically neutral, rather than theoretically loaded. Once we have found an NCC, one might hope that it will turn out to be a system dedicated to consciousness, or that it will turn out to yield an explanation of consciousness, but these are further questions. In the meantime the search for an NCC as defined poses a tractable empirical question with relatively clear parameters, one which researchers of widely different theoretical persuasions can engage in.

There are certain rewards of the search for an NCC that one might reasonably expect. For example, these systems might be used to monitor and predict the contents of consciousness in a range of novel situations. For example, we may be able to use them to help reach conclusions about conscious experience in patients under anesthesia, and in subjects with "locked-in syndrome" or in coma. In cases where brain architecture differs significantly from the original cases (perhaps some coma cases, infants, and animals), the evidence will be quite imperfect, but it will at least be suggestive.

These systems might also serve as a crucial step toward a full science of consciousness. Once we know which systems are NCCs, we can investigate the mechanisms by which they work, and how they produce various characteristic functional effects. Just as isolating the DNA basis of the gene helped explain many of the functional phenomena of life, isolating NCC systems may help explain many functional phenomena associated with consciousness. We might also systematize the relationship between NCCs and conscious states, and abstract general principles governing the relationship between them. In this way we might be led to a much greater theoretical understanding.

In the meantime, the search for a neural correlate of consciousness provides a project that is relatively tractable, clearly defined, and theoretically neutral, whose goal seems to be visible somewhere in the middle distance. Because of this, the search makes an appropriate centerpiece for a developing science of

consciousness, and an important springboard in the quest for a general theory of the relationship between physical processes and conscious experience.

References

- Anderson, R.A. 1997. Neural mechanisms in visual motion perception in primates. *Neuron* 18:865-872.
- Baars, B.J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Bogen, J.E. 1995. On the neurophysiology of consciousness, part I: An overview. *Consciousness and Cognition* 4:52-62.
- Bradley, D.C., Chang, G.C. & Andersen, R.A. 1998. Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature* 392:714-17.
- Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D.J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:200-19. Also in (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness* (MIT Press), and in (J. Shear, ed) *Explaining Consciousness: The Hard Problem* (MIT Press).
- Chalmers, D.J. 1998. On the search for the neural correlate of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Crick, F. and Koch, C. 1990. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2: 263-275.
- Crick, F. & Koch, C. 1995. Are we aware of neural activity in primary visual cortex? *Nature* 375: 121-23.
- Crick, F. & Koch, C. 1998. Consciousness and neuroscience. *Cerebral Cortex*.
- Dennett, D.C. 1991. *Consciousness Explain*. Little-Brown.
- Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.
- Eccles, J.C. 1994. *How the Self Controls its Brain*. New York: Springer-Verlag.
- Edelman, G.M. 1989. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.
- Engel, S., Zhang, X. & Wandell, B. 1997. Colour tuning in human visual cortex measured with functional

- magnetic resonance imaging. *Nature* 388 (6637):68-71.
- Flohr, H. 1995. Sensations and brain processes. *Behavioral Brain Research* 71:157-61.
- Gur, M. & Snodderly, D.M. 1997. A dissociation between brain activity and perception: Chromatically active cortical neurons signal chromatic activity that is not perceived.
- Güzeldere, G. 1999. There is no neural correlate of consciousness. Paper presented at "Toward a Science of Consciousness: Fundamental Approaches", Tokyo, May 25-28, 1999.
- He, S., Cavanagh, P. & Intriligator, J. 1996. Attentional resolution and the locus of visual awareness. *Nature* 384:334-37.
- Hobson, J.A. 1997. Consciousness as a state-dependent phenomenon. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Leopold, D.A. & Logothetis, N.K. 1996. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* 379: 549-553.
- Libet, B. 1982. Brain stimulation in the study of neuronal functions for conscious sensory experiences. *Human Neurobiology* 1:235-42.
- Llinas, R.R., Ribary, U., Joliot, M. & Wang, X.-J. 1994. Content and context in temporal thalamocortical binding. In (G. Buzsaki, R.R. Llinas, & W. Singer, eds.) *Temporal Coding in the Brain*. Berlin: Springer Verlag.
- Logothetis, N. & Schall, J. 1989. Neuronal correlates of subjective visual perception. *Science* 245:761-63.
- Marge, E. 1991. Magnetostimulation of vision: Direct noninvasive stimulation of the retina and the visual brain. *Optometry and Vision Science* 68:427-40.
- Merikle, P.M. & Reingold, E.M. 1992. Measuring unconscious processes. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.
- Milner, A.D. 1995. Cerebral correlates of visual awareness. *Neuropsychologia* 33:1117-30.
- Milner, A.D. & Goodale, M.A. 1995. *The Visual Brain in Action*. Oxford University Press.
- Newman, J.B. 1997. Putting the puzzle together: Toward a general theory of the neural correlates of consciousness. *Journal of Consciousness Studies* 4:47-66, 4:100-121.

- Penfield, W. 1937. The cerebral cortex and consciousness. In *The Harvey Lectures*. Reprinted in R.H. Wilkins (ed.) *Neurosurgical Classics* (New York: Johnson Reprint Corp., 1965).
- Penfield, W. & Rasmussen, T. 1950. *The Cerebral Cortex of Man: A Clinical Study of Localization of Function*.
- Pessoa, L., Thompson, E. & Noe, A. 1998. Finding out about filling in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences* 21:723-748.
- Salzman, C.D., Britten, K.H., & Newsome, W.T. 1990. Cortical microstimulation influences perceptual judgments of motion direction. *Nature* 346:174-77.
- Sheinberg, D.L. & Logothetis, N.K. 1997. The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences USA* 94:3408-3413.
- Shoemaker, S. 1981. Some varieties of functionalism. *Philosophical Topics* 12:93-119. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).
- Teller, D.Y. & Pugh, E.N. 1984. Linking propositions in color vision. In (J.D. Mollon & L.T. Sharpe, eds) *Color Vision: Physiology and Psychophysics*. London: Academic Press.
- Tong, F., Nakayama, K., Vaughan, J.T., & Kanwisher N. 1998. Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* 21:753-759.

First-Person Methods in the Science of Consciousness

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

Here are a few very general thoughts about how I see the shape of a science of consciousness, focusing on the issue of first-person methodology. At the end I will make a few remarks about how this might apply to the study of emotion.

As I see it, the science of consciousness is all about relating *third-person data* - about brain processes, behavior, environmental interaction, and the like - to *first-person data* about conscious experience. I take it for granted that there are first-person data. It's a manifest fact about our minds that there is something it is like to be us - that we have subjective experiences - and that these subjective experiences are quite different at different times. Our direct knowledge of subjective experiences stems from our first-person access to them. And subjective experiences are arguably the central data that we want a science of consciousness to explain.

I also take it that the first-person data can't be expressed wholly in terms of third-person data about brain processes and the like. There may be a deep connection between the two - a correlation or even an identity - but if there is, the connection will emerge through a lot of investigation, and can't be stipulated at the beginning of the day. That's to say, no purely third-person description of brain processes and behavior will express precisely the data we want to explain, though they may play a central role in the explanation. So *as data*, the first-person data are irreducible to third-person data.

The job of a science of consciousness, then, is to connect the first-person data to third-person data: perhaps to explain the former in terms of the latter, or at least to come up with systematic theoretical connections between the two. We ought at least to be able to come up with broad connecting principles, saying e.g. that certain sorts of experiences go along with certain sorts of processes in the brain (and/or vice versa), or that certain sorts of experiences go along with certain sorts of information-processing (and/or vice versa), and so on. If we're successful with this, perhaps we'll eventually be able to formulate simple and universal laws that underlie these broad connecting principles. That would be what I've called a "fundamental theory" of consciousness. We're a long way from that now, but we can at least make a start on connecting third-person data to first-person data at a broad level.

To do this, we need good methodologies for collecting the data and good languages and formalisms for expressing them. When it comes to the third-person data, these methods are very well-developed. Psychologists have developed sophisticated methods for studying behavior, for example, and neuroscientists have developed an ever-expanding group of ingenious methods for getting at what is going on in the brain: EEG, brain imaging, single-cell studies, and many others. And there are multiple formalisms for expressing these data: plain language, neurophysiological classification, various sorts of images and diagrams, computational models, and more. It seems fair to say that on the third-person side of things, the central constraints on data gathering and expression stem from technological (and ethical) limitations rather than conceptual barriers.

When it comes to first-person data, things aren't nearly so well-developed. Here methodologies for investigating the data are relatively thin on the ground, and formalisms for expressing them are even thinner. When it comes to methodologies, there have been various ideas: the 19th-century psychological introspectionists, 20th-century philosophical phenomenologists, and centuries of meditative studies in Eastern thought have all developed sophisticated frameworks, but all are widely held to have serious limitations, and none has been much integrated into contemporary science.

Contemporary scientists quite often do rely on first-person data in central ways, for example in psychophysics, where first-person experience of various phenomena such as illusions seems to be the coin of the realm in capturing the data that need to be explained. The methodology here seems to be that of simple untutored introspection and verbal report. This is not bad for capturing gross and simple features of conscious experience - does one see a pink splotch? - and maybe such methods will take us a fair way, but eventually we will need more to investigate that manifold intricacies of conscious experience.

When it comes to formalisms for expressing the first-person data, we are even worse off. Mostly we rely on simple language - an experience of red, of a horizontal line, a feeling of happiness, a sharp pain. But this sort of language is obviously coarse-grained and imprecise, and usually relies on an interlocutor's experience of the same phenomena to carry any communicative content at all. There have been a few attempts at developing more structured formalisms - the quantitative methods used in measurement of sensation in psychophysics, for example, or the structured phenomenal fields of Husserlian phenomenologists - but nothing with remotely the precision and scope of formalisms in the third-person domain.

In my opinion, the development of more sophisticated methodologies for investigating first-person data and of formalisms for expressing them is the greatest challenge now facing a science of consciousness. Only by developing such methodologies and formalisms will we be able to collect and express first-person data in such a way that it is on a par with third-person data, so that we can find truly systematic and detailed connections between the two.

When it comes to first-person methodologies, there are well-known obstacles: the lack of incorrigible access to our experience; the idea that introspecting an experience changes the experience; the impossibility of accessing all of our experience at once, and the consequent possibility of "grand

illusions"; and more. I don't have much that's new to say about these. I think that could end up posing principled limitations, but none provide in-principle barriers to at least initial development of methods for investigating the first-person data in clear cases. I hope to see ideas from Western and Eastern philosophy and from contemporary and historical psychology integrated with a series of new ideas in coming years.

When it comes to first-person formalisms, there may be even greater obstacles: can the content of experience be wholly captured in language, or in any other formalism, at all? Many have argued that at least some experiences are "ineffable". And if one has not had a given experience, can any description be meaningful to one? Here again, I think at least some progress ought to be possible. We ought at least to be able to develop formalisms for capturing the *structure* of experience: similarities and differences between experiences of related sorts, for examples, and the detailed structure of something like a visual field. I don't know what exactly such a formalism would look like, but perhaps something bringing in ideas from geometry or topology, or from information theory, might be useful.

As for the intrinsic non-structural aspects of experience (the sensation of red, for example), things are more difficult. But even here one could arguably find some underlying structure: e.g. color experiences can arguably be decomposed into experiences of brightness, saturation, and hue. Perhaps - let's speculate - one might develop a theory of "proto-qualia" from which the qualia we experience are systematically built up? Or perhaps not, in which case we'll need other ideas. The idea of simple building blocks might help to some extent with the problem of communication, though: although different individuals may have different experiences, arguably some of the same building blocks might be present in each case. So perhaps they could abstract the primitive elements through inference from their own experience, and then get some idea of others' experience through the idea of recombination. Perhaps this could even eventually (when connected appropriately to third-person data) give us some clue about the subjective lives of animals. Or again, perhaps not.

What about emotions, in particular? Here I don't have much to say, and I expect that other participants in this symposium have thought about the issue in far more sophisticated ways than I have. But I hope I will be forgiven for entering into the spirit of things with a little uninformed speculation.

It's clear at a glance that when it comes to first-person study of emotion, the issues of both methodology and formalism are relevant. How does one collect first-person data about emotional experience? There are presumably particular difficulties with reliability here. How reliable can one expect an observer in a red-hot rage to be? In the domain of emotion, isn't self-deception likely to be ubiquitous? And presumably there will be observer effects all over the place: it doesn't seem implausible that cultivating a detached perspective on emotional experience will change the character of the experience significantly.

On the positive side, many people seem to be quite good at investigating their own emotional states, and it is a particularly interesting project. In this area, going beyond gross features to subtleties may be particularly rewarding. This is illustrated in the rich investigations of novelists such as Proust. Perhaps there is some way to tap into this sort of thing for scientific purposes?

I don't know much about the field, but my guess is that right now, the dominant methods for accessing

first-person data in scientific experiments on emotion involve relatively untutored introspection of relatively gross features: asking a subject whether they are having experiences of happiness and sadness, and the like. And I imagine that even this provides a productive source of data to be going on with, and with which a lot of interesting science can be done. I imagine that participants in this symposium will be talking about some more developed methodologies for the first-person study of emotion, though, and I will be interested to hear what they have to say.

As for formalisms, this seems to be more of a question mark. Emotions seem to be particularly inexpressible, especially to one who hasn't experienced the emotion in question before. Even where two individuals have emotional experience in common, it can be hard to find the right language to describe it. At the same time, our experiences clearly vary on a number of clear dimensions: duration, intensity, positive or negative affect, and numerous others. And I imagine that most of these things are already exploited by experimenters in the field. It's far from clear (to me, at least) just how much of the complex character of emotions can be captured in such quantitative and structural measures, but it's at least a start. And perhaps we will be able to develop more and more sophisticated formalisms for expressing more and more of the complex structure of emotion, so the unexpressed residue will at least shrink considerably. I imagine there are a good number of ideas along these lines out there already.

As to what to do with that unexpressed residue: perhaps we'll have to rely on common language to bootstrap our understanding of common elements of experience, or perhaps we'll be able to go further with some sort of building-block methodology. Or perhaps we'll come to the conclusion that formalisms can only tell us so much about emotions, and that novelists are needed to tell us the rest. I don't have any firm expectations here myself, but I'll be very interested to see how things play out, both in this conference and in coming years.

Insentience, Indexicality, and Intensions

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

*[[This is a commentary on John Perry's book *Knowledge, Possibility, and Consciousness*. It is destined for a symposium on Perry's book in *Philosophy and Phenomenological Research*.]]

John Perry's book *Knowledge, Possibility, and Consciousness* is a lucid and engaging defense of a physicalist view of consciousness against various anti-physicalist arguments. It will come as no surprise to Perry to learn that I think his defense does not succeed. In what follows, I will address Perry's responses to the three main anti-physicalist arguments he discusses: the zombie argument (focusing on insentience), the knowledge argument (focusing on indexicality), and the modal argument (focusing on intensions).

1 The Zombie Argument

Zombies are unconscious creatures that are microphysically identical to conscious beings such as ourselves. If zombies are metaphysically possible, then materialism is false. The zombie argument aims to establish that zombies are conceivable, infers that they are metaphysically possible, and concludes that materialism is false.

Perry's response is straightforward. He says (pp. 77-80) that zombies will be possible if and only if epiphenomenalism is true. And he says that he cannot see any reason why one would think that zombies (and zombie worlds) are possible unless one were *already* an epiphenomenalist. So the zombie argument begs the question, by tacitly assuming epiphenomenalism as a premise.

I think that Perry is clearly wrong here. Whatever the merits of the zombie argument, it does not beg the question. To see this, note that the zombie argument is not based on a *single* premise - the possibility of zombies - but rather on two premises. The first premise is that zombies are conceivable, roughly in the sense that there is no a priori contradiction in the idea of a zombie. The second premise is that if zombies are conceivable in this sense, then they are possible. (Or they are "primarily possible", in that there is a possible world satisfying the relevant primary intension).

Each of these premises is accepted by many philosophers who clearly reject epiphenomenalism. For

example, the first premise is accepted by many "type-B" materialists (those who accept an epistemic gap between the physical and the phenomenal but deny an ontological gap), all of whom deny epiphenomenalism: e.g., Ned Block, Chris Hill, Joe Levine, Brian Loar, and many others. The second premise is accepted by many "type-A" materialists (those who deny an epistemic gap between physical and phenomenal) and interactionists, all of whom deny epiphenomenalism: e.g., Frank Jackson (current incarnation), David Lewis, Richard Swinburne, and others. Of course few or none of these people accept both premises. But their mere existence shows that the premises have support that is independent of a prior acceptance of epiphenomenalism.

To go beyond a roll-call, it can be seen straightforwardly that each premise has substantive support that goes beyond epiphenomenalism. The second premise rests on general considerations linking conceivability and possibility. The first premise rests partly on *prima facie* conceivability intuitions that many share, and partly on deeper considerations concerning the absence of any conceptual linkage between microphysical concepts (which are structural-functional in nature) and phenomenal concepts (which are not). In both cases, whether or not these premises are correct, their support presupposes nothing about epiphenomenalism.

I suppose that Perry might try to argue that the first premise tacitly builds in epiphenomenalism. Given all the non-epiphenomenalists who accept the premise, however, this would seem unlikely: it would require that they be deeply irrational, or have deeply divided minds. Further, Perry says a number of things elsewhere in the book that seem to come close to embracing the first premise (in the relevant sense of conceivability). For example, although he embraces an identity between phenomenal and physical properties, he denies (pp. 87-88) that consciousness supervenes logically on the microphysical (i.e., he denies a conceptual entailment from microphysical truths to phenomenal truths), which suggests that there will be no *a priori* contradiction in the notion of a zombie. And even in raising questions about zombies, he points to nothing like an incoherence in the notion; at best he points to a tension with certain *a posteriori* beliefs about the causal role of our own conscious states.

In any case, it seems clear that to reject the argument, Perry has to go beyond considerations about begging the question, and instead give substantive reasons to reject one of the premises. From Perry's overall position, I would expect him to be more likely to reject the second premise than the first. Like other type-B materialists, he can accept that zombies are conceivable in the relevant sense, but deny the principle linking conceivability (in this sense) with possibility. I will return to this matter in the last section.

One final point: it is worth noting that even the possibility of zombies does not obviously entail epiphenomenalism. To see this, note that an interactionist dualist can accept the possibility of zombies, by accepting the possibility of physically identical worlds in which physical causal gaps (those filled in the actual world by mental processes) go unfilled, or are filled by something other than mental processes. The first possibility would have many unexplained physical events, but there is nothing metaphysically impossible about unexplained physical events. Also: a Russellian "panprotopsyichist", who holds that consciousness is constituted by the unknown intrinsic categorical bases of microphysical dispositions, can accept the possibility of zombies by accepting the possibility of worlds in which the microphysical

dispositions have a different categorical basis, or none at all. So even if the argument had the one-premise structure that Perry suggests, then as long as it was appropriately neutral on the physical character of the world, it would not obviously beg the question in favor of epiphenomenalism.

2 The Knowledge Argument

The knowledge argument holds that there is phenomenal knowledge (e.g. concerning what it is like to see red) that is not deducible from physical knowledge, and infers that phenomenal facts are non-physical facts. Perry's response, essentially, is to analyze phenomenal knowledge as a sort of indexical knowledge. We know that there is indexical knowledge (e.g. concerning my current location) that is not deducible from complete physical knowledge, but this indexical knowledge does nothing to falsify physicalism. If phenomenal knowledge is indexical knowledge, then we can straightforwardly explain the epistemic gap between the physical and phenomenal domains, without requiring any non-physical ontology.

Perry applies this strategy to the new knowledge that (physically omniscient) Mary gains when she sees red for the first time, roughly as follows. Mary gains new knowledge of the form "red things cause experiences of type *P*". This knowledge crucially involves a phenomenal concept *P* - a concept of what it is like to have a certain sort of experience. Perry's strategy (pp. 145-48) is to analyze Mary's phenomenal concept *P* as a demonstrative concept - which he labels *this_i* - that functions, like other demonstrative concepts, to pick out whatever sort of experience she is currently attending to. Because of an underlying indexical character, knowledge involving demonstrative concepts cannot usually be inferred from complete objective knowledge. So Mary's epistemic gap is reduced to the familiar epistemic gap for indexicals and demonstratives.

My response is straightforward. I think that there are clear reasons to hold that phenomenal knowledge (of the crucial sort) is *not* indexical knowledge, and that phenomenal concepts (of the crucial sort) are *not* demonstrative concepts of this sort; so this analysis of the epistemic gap fails. Mary does gain indexical knowledge involving a demonstrative concept, but this indexical knowledge is not the phenomenal knowledge that is central to the knowledge argument, and this demonstrative concept is not the concept that Mary's central new knowledge involves. I have argued for this based on an independently-grounded analysis of phenomenal concepts elsewhere (Chalmers 2002a), but I will give some reasons here.

It is useful to consider analogies with other demonstrative knowledge of types. Let *this_S* be a demonstrative concept of certain *shapes*. Jill might tell Jack that she is about to show him her favorite shape. When she shows him a circle, he might form the thought "Jill's favorite shape is *this_S*", where *this_S* refers to circles. This is a clearly demonstrative thought. He might also form the thought "Jill's favorite shape is a circle". This is a clearly nondemonstrative thought: the right hand side is not a demonstrative concept, but what we might call a *qualitative* concept. Finally, he might form the thought "*this_S* is a circle". This is a substantive, nontrivial thought, taking the form of an identity involving a demonstrative concept and a qualitative concept. This sort of thought is very common with demonstratives - one conceives the object of a demonstration *as* the object of a demonstration, and at the same time attributes it substantive qualitative properties, conceived non-demonstratively.

On the face of it, Mary's situation is precisely analogous. She knows that she is about to have an experience of the sort usually caused by red things. Upon having the experience, she might form the thought "The experience usually caused by red things is *this_I*". But at the same time, she might form the thought "The experience usually caused by red things is *R*", where *R* is a *qualitative* concept of the sort of experience she is now having. This is brought out by the fact that just as Jack might think the substantive thought "this_S is a circle", Mary might think the substantive thought "this_I is *R*". Like Jack's thought, Mary's thought involves attributing a certain substantive qualitative nature to an object that is identified demonstratively. The concept *R* - her qualitative concept of the sort of experience in question - is not a demonstrative concept at all, as witnessed by the nontriviality of the identity involving the demonstrative concept *this_I* and *R*. And Mary's *crucial* new knowledge in the thought-experiment, the new knowledge that the knowledge argument turns on, is the substantive qualitative knowledge involving *R* (e.g. that red things cause *R* experiences), not the relatively uninteresting indexical knowledge involving *this_I*.

One might be distracted by the fact that Mary's concept *R* is a *new* concept, acquired upon having the experience, while Jack's concept "circle" is an old concept. But this is inessential to Jack's case. We can imagine that Jack has never seen a circle before (if one is worried about plausibility, make it another shape), but that on seeing a circle for the first time, he acquires the qualitative concept of circularity. He will then be in the position to think the qualitative thought "Jill's favorite shape is a circle", and to think the substantive demonstrative-qualitative thought "This_S is a circle. Mary's situation is just the same. Upon seeing red for the first time, she is able to form a qualitative concept of that sort of experience (based on her exposure to it), to think qualitative thoughts involving that concept, and to think demonstrative-qualitative thoughts in which both a demonstrative and a qualitative concept are deployed.

Perry might respond by holding that the cases are deeply disanalogous, and that where Joe has two concepts ("this_S" and "circle"), Mary has only one (*this_I*). But this seems quite false to the phenomenology of the cases, both of which involve substantive knowledge of the form "this is such-and-such", where "such-and-such" is a qualitative predication based on exposure to a demonstrated entity. We can also note that just as it seems that Joe could have had different thoughts of the form "this_S is a square", if he had been exposed to a different shape, it seems that Mary could have had different thoughts of the form "this_I is *G*" had she been exposed to a different sort of experience, such as an experience of green. So it seems that the qualitative concept of the experiential type is quite different from the demonstrative concept.

I conclude that Perry gives an adequate analysis of certain subsidiary demonstrative knowledge that Mary gains - knowledge that is analogous to demonstrative knowledge in other domains. But Perry gives no adequate analysis of Mary's substantive new non-indexical, non-demonstrative phenomenal knowledge, and in particular gives no way of reconciling this new phenomenal knowledge with the truth of physicalism.

(In Chalmers (2002a), I distinguish three sorts of phenomenal concepts: pure phenomenal concepts (such as *R*), demonstrative phenomenal concepts (such as *this_I*) and relational phenomenal concepts (such as

"the sort of experience typically caused by red things"). Perry's discussion seems to acknowledge only two sorts, the demonstrative and the relational; or at least, it seems to assimilate the pure phenomenal and the demonstrative with each other.)

One can also make a direct case against any analysis of phenomenal knowledge as indexical or demonstrative knowledge, as follows. In the indexical case, any epistemic gaps disappear from an objective perspective. Say that I am physically omniscient, but do not know whether I am in the USA or Australia (let's imagine that there are appropriate qualitative twins in both). Then I have a certain indexical ignorance, and discovering that I am in the USA will constitute new knowledge. But if someone *else* is watching from the third-person point of view and is also physically omniscient, they will have no corresponding ignorance: they will know that A is in Australia and that B is in the US, and that's that. There is no potential knowledge that they lack: from their perspective, they know everything there is to know about my situation. So my ignorance is *essentially* indexical, and evaporates from the objective viewpoint. The same goes for indexical ignorance concerning what time it is, for demonstrative ignorance concerning what *this* is, and so on. In all these cases, the ignorance disappears from the objective viewpoint: an objectively omniscient observer can know everything there is for them to know about my situation, and there will be no doubts for them to settle.

Now consider Mary's ignorance. From her black-and-white room, she is ignorant of all sorts of facts: what it will be like for her to see red for the first time, what it is like for others to see red, and so on. Only the first of these looks even apparently indexical, so let us focus on that. In this case, a physically omniscient observer may have precisely analogous ignorance: even given his complete physical knowledge, he may have no idea what it will be like for Mary to see red for the first time. So this ignorance does not evaporate from the objective viewpoint. The same goes even more strongly for knowledge of what it is like for others to see red. For any observer, regardless of their viewpoint, there will be an epistemic gap between complete physical knowledge and this sort of phenomenal knowledge. This suggests very strongly that phenomenal knowledge is not a variety of indexical or demonstrative knowledge at all. Rather, it is a sort of objective knowledge of the world, not essentially tied to any viewpoint.

If this is right, then any analysis of phenomenal concepts as indexical or demonstrative concepts fails, and any attempt to explain Mary's epistemic gap in terms of the epistemic gap for indexical or demonstrative concepts fails. If this is right, then Perry's substantive account of Mary's epistemic gap, in terms of the "reflexive content" of certain beliefs, will also fail. Reflexive content is essentially indexical content (tied to the relation of *this* token to the world), so that an analysis in these terms applies only to broadly indexical knowledge.

A similar diagnosis can be applied to analyses of phenomenal knowledge by many other materialists who seek to acknowledge Mary's epistemic gap and to explain it away: any analysis of phenomenal concepts as indexical, demonstrative, or recognitional concepts will fall prey to similar objections. This does not show by itself that no analysis of phenomenal knowledge can be given that saves both physicalism and the epistemic gap but I think that it renders the prospects dim.

3 The modal argument

The third anti-physicalist discussion that Perry discusses is the "modal argument". Of course the zombie argument is itself a sort of modal argument. Presumably Perry intends a more general target here, but I think some of the same issues arise. The chapter is divided into a discussion of "Kripke's argument" and of "Chalmers' argument". His discussion of the latter focuses on my use of the two-dimensional framework in arguing against materialism; so this can be seen as filling in the issues that went unaddressed in his earlier discussion, concerning the relationship between conceivability and possibility. I will focus on this issue here.

Perry's discussion in this chapter is harder to follow than in the other chapters of the book, and it is not easy to clearly identify his responses to the arguments in question. His basic strategy appears to be to agree with his opponents that where there is "apparent contingency" (Kripke) or "conceivability" (me) of some mental-physical dissociation, there is some sort of possibility in the vicinity, but to argue that once the possibility is appropriately understood (in terms of reflexive relations to tokens and the like), it does not threaten materialism. I was left unclear, however, on how this strategy bears on the specific argument I put forward.

To summarize the argument briefly (see Chalmers 2002b for a more detailed treatment): Let us say that S is conceivable when it is not a priori that $\sim S$. Any given statement S can be associated with two intensions: a *primary intension*, which is a function from centered worlds (worlds plus individuals/times) to truth-values, and a *secondary intension* which is a function from uncentered worlds to truth-values. When S is necessary, it has a necessary secondary intension (and vice versa); when S is a priori, it has a necessary primary intension (and somewhat controversially, vice versa). Let us say that S is *primarily possible* when its primary intension is true at some centered world. Let P be the complete microphysical truth about the world. Let Q be a phenomenal truth. Then the anti-materialist argument can be put as follows

- (1) $P \& \sim Q$ is conceivable
- (2) If S is conceivable, S is primarily possible.
- (3) If $P \& \sim Q$ is primarily possible, materialism is false.
-
- (4) Materialism is false.

Here, the first premise is simply a statement of the epistemic gap, grounded in the a priori coherence of zombies, or in the a priori coherence of inverted spectra, or in Mary's inability to deduce phenomenal truths from physical truths. The second premise is a core principle of the two-dimensional framework,

linking apriority and primary possibility. The third premise is not completely obvious (since materialism requires the *secondary* possibility of $P \& \sim Q$), but some relatively straightforward argument involving the nature of physical and phenomenal concepts takes one from the relevant primary possibility to a relevant secondary possibility.

(On the third premise: If a world satisfies the primary intension of P , it is at least structurally identical to our world. So either (i) the world is completely physically identical to our world but different in mental respects, so materialism is false, or (ii) it differs at most in the intrinsic microphysical properties that underlie this structure. Exploiting (ii) leads directly to the Russellian "panprotopsychoist" position, which I count for present purposes as a version of nonmaterialism. The argument also needs supplementing to handle the role of indexicals and centering, but this is not hard: one can argue that premise 1 is true even when one supplements P with full indexical "locating" knowledge (see below for more on this), and then a corresponding version of premise (3) still goes through.)

Despite Perry's extensive discussion of the two-dimensional framework, I was left unclear on which premise of this argument he would reject, and why. For reasons discussed earlier, he seems to be committed to a version of premise 1. Many of Perry's fellow type-B materialists reject premise 2 (often allowing that it applies in all standard cases, but holding that the case of consciousness is exceptional or unique). Perhaps Perry would reject this premise, but he does not discuss it explicitly, so it is hard to tell. Perhaps more likely, given Perry's sympathy for the idea that where there is apparent contingency, there is possibility, it is not out of the question that he accepts the second premise and rejects the third, holding that the relevant possibility does not trouble materialism. But again, it is hard to tell from Perry's discussion, as he does not address the argument directly, and he does not argue against the relevant premises.

Insofar as I understand Perry's response to the two-dimensional argument (on pp. 198-200), it focuses on the status of an identity such as " $Q_R = B_47$ ", where Q_R is the concept of whatever sort of experience is caused by red things, and B_47 is the concept of brain states of a certain sort. He allows that the identity is "apparently contingent" (conceivable in the relevant sense), allows that there is a primary possibility in the vicinity, but explains the relevant primary possibility in terms of the fact that (i) Q_R might have picked out a different phenomenal property, and (ii) B_47 might have picked out a different sensation. In two-dimensional terms (modulo a few complications in the translation that I pass over here), it seems that Perry is accepting: (i) the primary intension of Q_R picks out a different phenomenal property in some centered worlds, and (ii) the primary intension of B_47 picks out a different sensation in some centered world. In effect, this is to explain the primary possibility in terms of the fact that Q_R and B_47 have distinct primary and secondary intensions. As such, it appears that Perry is here allowing premise (2) above, but denying the equivalent of premise (3): there is no argument from the relevant primary possibility to the falsity of materialism.

There are a few things to say here. First, my own argument here is not formulated in terms of identities, precisely because it is hard to mount an argument from the primary possibility of the negation of a physical-phenomenal identity to the falsity of materialism. (Perry characterizes my argument in terms of identity on pp. 188-89, but this is a mischaracterization.) When we invoke the full claim $P \& \sim Q$, rather

than the mere negation of an identity, the argument from primary possibility to the falsity of materialism is much more straightforward. So Perry is addressing something of a straw figure here.

Second, Perry invocation of the *relational* phenomenal concept Q_R is very odd in this context. This is a deeply extrinsic characterization of the phenomenal property in question, with a very different primary and secondary intension. For the anti-materialist argument it is much better to invoke something closer to a pure phenomenal concept, to accommodate Kripke's insight that for phenomenal concepts, there is no gap between reference-fixers and reference (or between primary and secondary intensions). Perhaps Perry denies that there are any such phenomenal concepts; but if so, he needs to give substantive arguments against the Kripkean claim. In any case, it is clear that by appealing to the relational concept, Perry is making things easy for himself. (At the very least, I would have expected to see the demonstrative concept *this_I*.) If there are phenomenal concepts with the same primary and secondary intension, then Perry's strategy here will fail. (In fact, even if phenomenal concepts have different primary and secondary intensions, the anti-materialist argument still goes through, but things are more complicated.)

Third, Perry appeals to the possibility that there are centered worlds in which (the primary intension of) B_47 picks out a different sensation. It is hard to see how this is supposed to go, at least if B_47 gives a full microphysical description of the relevant state. As such, it seems that there is no centered world in which B_47 picks out a different *physical property*. And presumably Perry holds that there are no centered worlds in which that very physical property is a different sensation (that would deny the necessity of identity). One way out might be to endorse the Russellian claim that physical concepts are topic-neutral, fixing reference to different underlying intrinsic properties depending on the state of the world. But this leads only to the "panprotopsychist" loophole, which I don't think Perry is exploiting. So it is hard to see how this move works for him.

I would have expected Perry to embrace an alternative strategy, suggested by his response to the knowledge argument. If A is "I am in the USA", then $P \& \sim A$ is conceivable and primarily possible, but the primary possibility simply corresponds to the actual world with a different center. Such a centered world yields no argument against materialism. By analogy, Perry might hold that $P \& \sim Q$ is conceivable because of a tacit indexicality in Q , and is primarily possible, but the primary possibility simply corresponds to the actual world with a different center, again yielding no argument against materialism. Here the center would presumably need to be expanded beyond an individual and a time, but that could be justified if phenomenal concepts were truly analogous indexical concepts, so that phenomenal information is a sort of "locating information". For example, one might supplement the "you are here" and "now is here" arrows at the center with "this is here" arrows pointing to certain states as the referents of phenomenal demonstratives. Then the relevant centered world will simply be a world with different locations for the "this is here" arrows.

I think that this move does not work, for more or less the reasons given above. First, phenomenal concepts are not indexical or demonstrative concepts, for reasons given there. Second, phenomenal information does not disappear from the third-person viewpoint, so it is not locating information (it is epistemically objective information), and so not the sort of thing that can be built into the center of a

world. Third, while Mary is in the black and white room, one can supplement her physical information with all the locating information one likes - "you are here", "now is here", and "this is here" with arrows pointing to various of her brain states - and she will still be in complete ignorance of what it is like to see red. So unlike the case of my physical location, the epistemic gap here is not closed by adding any amount of locating information. Still, this is an interesting strategy that is worth exploring, especially in the context of Perry's response to the knowledge argument.

In any case: it is possible that I have misunderstood Perry's intended response to the modal argument here. So I would be interested to see how he responds to the straightforward three-premise argument above.

4 Conclusion

Perry's responses to the anti-materialist arguments are philosophically deep and sophisticated. Nevertheless, I think that upon close examination, these responses fail in straightforward ways. Given Perry's depth and sophistication, I take this to be indirect evidence that the arguments are sound.

References

Chalmers, D.J. 2002a. The content and epistemology of phenomenal belief. In (A. Jokic and Q. Smith, eds) *Aspects of Consciousness*. Oxford University Press.
[www.u.arizona.edu/~chalmers/papers/belief.html]

Chalmers, D.J. 2002b. Does conceivability entail possibility? In (T. Gendler and J. Hawthorne, eds) *Imagination, Conceivability, and Possibility*. Oxford University Press.
[www.u.arizona.edu/~chalmers/papers/conceivability.html]

Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. MIT Press.

Availability: The Cognitive Basis of Experience?

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

[[This was written as a commentary on [Ned Block's](#) paper "[On A Confusion about a Function of Consciousness](#)". It appeared in [Behavioral and Brain Sciences](#) 20:148-9, 1997, and also in the collection *The Nature of Consciousness: Philosophical Debates* (MIT Press, 1997) edited by Block, Flanagan, and Guzeldere.]]

Introduction

Block's distinction between access consciousness and phenomenal consciousness (or experience) is very useful. There is clearly a *conceptual* distinction here, as illustrated by the facts that: (1) one can *imagine* access without experience and vice versa; (2) access can be observed straightforwardly, whereas experience cannot; and, most important, (3) access consciousness seems clearly amenable to cognitive explanation, whereas phenomenal consciousness is quite perplexing in this regard. But the tight *empirical* link between the two phenomena deserves attention.

Bringing access and experience closer together.

Block himself notes that P-consciousness and A-consciousness often occur together. This is no accident, as one can see by noting that a P-conscious experience is usually reportable, and that reportability implies accessibility of the corresponding information. Block does not think they always occur together, but I think that with appropriate modifications they might. One of the most interesting projects in this area is that of *modifying* the concept of A-consciousness in such a way as to make it plausible that A-consciousness (in the modified sense) and P-consciousness are perfect correlates.

A good start is the modified notion of *direct availability for global control*. That is, a content is A-conscious in the modified sense when it is directly available for use in directing a wide range of behaviors, especially deliberate behaviors. I am not sure how different this is from Block's definition: it plays down the role of rationality and reasoning (after all, impairments of rationality probably do not diminish phenomenal consciousness), it relegates verbal report to the status of a heuristic (as Block himself suggests), and there is another important difference that I will come to shortly. The restriction to *direct* availability works to eliminate contents that can be retrieved with some work but that are not conscious.

To see how well this modified notion of A-consciousness correlates with P-consciousness, we need to see how it handles Block's examples in which one sort of consciousness occurs without the other. Block's examples of A-consciousness without P-consciousness are all mere conceptual possibilities (zombies and super-blindsight, for example), so they are not relevant here, but to illustrate P-consciousness with A-consciousness he gives some real-world examples. One is Sperling's example in which all nine letters in a square array are experienced, but only three can be reported at a time. In this case, only three letter-representations are *accessed*, but it is nevertheless plausible that each of the nine was *available*, until the process of access destroyed their availability. This works because the modified notion of A-consciousness is *dispositional* - not access but accessibility is required. And it is plausible that all nine letter-representations are A-conscious in the modified sense. So even in this case, P-consciousness and modified A-consciousness occur together.

The case of the drilling noise in the background can be handled similarly. Here it seems reasonable to say that the information was directly available all along; it simply wasn't accessed. The case of experience under anesthesia (if this is actual) is trickier, but we might handle it by saying that in these cases the corresponding contents are *available* for global control; it is just that the control mechanisms themselves are mostly shut down. We might say that the information makes it to a location where it could have been used to direct behavior, had the motor cortex and other processes been functioning normally.

Other cases could be considered and further refinements could be made. A fuller account might flesh out the *kind* of availability required (perhaps a kind of high-bandwidth availability is required for experience, or at least for experience of any intensity) and might specify the relevant kind of control role more fully. Counterexamples are not threatening but helpful; they allow us to refine the definition further. The details can be left aside here; the point is that this project will lead to a functionally characterized property that might correlate perfectly with P-consciousness, at least in the cases with which we are familiar.

This property - something in the vicinity of direct availability for global control - could then be thought of as the information-processing correlate of P-consciousness, or as the cognitive basis of experience. There are some interesting consequences for the issues that Block discusses.

Empirical work on consciousness.

Block notes that researchers on consciousness often start with an invocation of phenomenal consciousness but end up offering an explanation of A-consciousness and leaving P-consciousness to the side. The tight link between the two suggests that a somewhat more charitable interpretation is possible. If experience correlates with availability for global control, much of this work can be interpreted as seeking to *explain* A-consciousness, but trying to find a *basis* for P-consciousness. For example, Crick and Koch's oscillations are put forward because of a potential role in binding and working memory; that is, in integrating contents and making them available for control (working memory is itself an availability system, after all). If both the empirical hypothesis (oscillations subserve availability) and the bridging

principle (availability goes along with experience) are correct, then the oscillations are a neural correlate of experience, which is just what Crick and Koch claim.

The same holds elsewhere. Shallice's "selector inputs" for "action systems" (1972) and his "Supervisory System" (1988a; 1988b) are clearly supposed to play a central role in availability and control; if the empirical hypothesis is correct, these could reasonably be regarded as part of the basis for conscious experience. Similarly, the "global workspace" of Baars (1988), the "high-quality representations" of Farah (1994), the "temporally-extended neural activity" of Libet (1993), and many other proposals can be all be seen as offering mechanisms in the process whereby some contents are made available for global control. The common element is striking. Of course, it is an empirical question which of these proposals is correct (although more than one might be, if they offer accounts of different parts of the process or descriptions at different levels). But insofar as these mechanisms play a role in the availability/control process, they are candidates to be neural or cognitive correlates of experience, which is often what the authors suggest (correlation is all that Farah and Libet claim; Shallice and Baars oscillate between "correspondence" and explanation).

The picture is this: (1) we know that availability goes along with experience; (2) we empirically discover that some mechanism plays the central role in the availability process. We may then conclude that the mechanism is part of the explanation of A-consciousness and part of the basis of P-consciousness. Of course, the story about the mechanism does not alone *explain* P-consciousness, as we still have not explained why availability always goes along with experience; we have simply taken for granted that it does. But if we are prepared to take the link between availability and experience as a kind of background assumption (perhaps for later explanation), this can provide a useful partial explanation of the contents of experience.

A phenomenal consciousness module?

Interestingly, this analysis allows us to make some sense of the idea of a phenomenal consciousness module. *If* it turns out that there is a single system responsible for mediating the availability of certain contents for global control - something like Baars' global workspace or Shallice's supervisory system - then it might be plausible that the contents of that system correspond precisely to the contents of experience, and maybe we could call it a P-consciousness module. I do not think it is probable that there is such a module - more likely there are many different mechanisms by which contents become available for a control role - but at least the idea makes sense. But the *only* way there could be a "P-consciousness" module would be for it to be an availability/control module. If a module were dissociable from the relevant role in availability and control, the considerations above suggest that it would be dissociable from P-consciousness too.

In particular, there is something very strange about the idea of an "epiphenomenal" P-consciousness module (Block's Figure 3). The main motivation for epiphenomenalism is surely that experience seems superfluous to any information-processing; but Block's idea suggests an implausible epiphenomenalism *within* the information-processing story. Indeed, if the module has no effect on other processes, then we

could lesion it with no external change (same reports, even), and no empirical evidence could support the hypothesis. Perhaps Block means to allow that the module has the very limited function of causing phenomenal reports, so that lesioning it eliminates remarks such as "I am having a blue sensation." But now either (1) remarks such as "There is a blue object", confident blue-directed behavior, and so on are all eliminated too - in which case the module had an important function after all - or (2) they are preserved (a kind of ultra-superblindsight), implying an extraordinary independence between the pathways responsible for phenomenal report and those responsible for visual descriptions and normal visual processing. Given the remarkable coherence between visual descriptions and reports of visual experience, one presumes that they are tied more closely than this.

The function of consciousness?

The link between P-consciousness and (modified) A-consciousness makes the search for a function for P-consciousness even more hopeless. Given the correlation, *any* purported function for P-consciousness can be attributed to A-consciousness instead.

Only those who implausibly identify the *concept* of P-consciousness with that of (modified) A-consciousness have a way out. If one accepts the conceptual distinction, one will accept the conceivability of zombie functional isomorphs (made of silicon, say). To be consistent, one must then accept the conceivability of zombie *physical* isomorphs, as there is no more of a conceptual entailment from neural stuff to consciousness than there is from silicon stuff. From here, it is easy to see that P-consciousness gives me no functional advantage. After all, I am different from my zombie twin *only* in that I have P-consciousness and he does not, but we are functionally identical.

Block suggests that P-consciousness might "grease the wheels" of A-consciousness, but this cannot work. P-consciousness is redundant to the explanation of the physical mechanisms of A-consciousness, as the conceivability of the zombie shows: same physical mechanisms, same explanation of A-consciousness, no P-consciousness. The remaining option is to "identify" P-consciousness with modified A-consciousness (empirically but not conceptually), solving the problem by fiat. I think this sort of identification without explanation misunderstands the way that scientific identification works (see Chalmers 1995; Jackson 1994), but in any case it still leaves the concept of P-consciousness with no *explanatory* role in cognitive functioning. The independent concept of A-consciousness does all the work. I think it best to accept instead that phenomenal consciousness is distinct from any physical or functional property, and that it does not need to have a function to be central to our mental lives.

References

Baars, B.J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Crick, F. and Koch, C. 1990. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2: 263-275.

Farah, M.J. 1994. Visual perception and visual awareness after brain damage: A tutorial overview. In (C. Umiltà and M. Moscovitch, eds.) *Consciousness and Unconscious Information Processing: Attention and Performance* 15. MIT Press.

Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, and S. White, eds.) *Philosophy and the Cognitive Sciences*.

Shallice, T. 1972. Dual functions of consciousness. *Psychological Review* 79:383-93.

Shallice, T. 1988a. *From Neuropsychology to Mental Structure*. Cambridge University Press.

Shallice, T. 1988b. Information-processing models of consciousness: possibilities and problems. In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science*. Oxford University Press.

Self-Ascription Without Qualia: A Case-Study

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

[[Published in *Behavioral and Brain Sciences* 16:35-36, 1993, as a commentary on Alvin Goldman, [The Psychology of Folk Psychology](#).]]

In Section 5 of his interesting article, Goldman suggests that the consideration of imaginary cases can be valuable in the analysis of our psychological concepts. In particular, he argues that we can imagine a system that is isomorphic to us under any functional description, but which lacks qualitative mental states, such as pains and color sensations. Whether or not such a being is empirically possible, it certainly seems to be logically possible, or conceptually coherent. Goldman argues from this possibility to the conclusion that our concepts of qualitative mental states cannot be analyzed entirely in functional terms.

This thought-experimental methodology seems sound to me, and I agree with Goldman on the logical possibility of these absent-qualia cases (although many functionalists would not; e.g. Armstrong 1968, Dennett 1991, Shoemaker 1975). However, I think that this methodology can be taken further, yielding conclusions that oppose those that Goldman draws elsewhere in the article.

Consider: if it is logically possible that my functional isomorph might lack qualia entirely, it seems equally logically possible that there could be a qualia-free *physical replica* of me. We have already seen that there is no conceptual entailment relation from the functional properties of a system to the qualitative properties; it seems even clearer that there is no entailment relation from the non-functional implementational details to qualia. (What conceptual entailment could neurophysiological detail possibly provide that silicon, or even Chinese nations, could not?) So let's consider **Zombie Dave**, my qualia-free physical replica. Zombie Dave is almost certainly not an empirical impossibility, but he is a conceptual possibility.

First, let us ask: Does Zombie Dave have beliefs? It seems to me that he does. If we ask him where his car is, he'll tell us that it's in the driveway. If we ask him whether he likes basketball, he tells us that he does. If we tell him that there's a basketball game starting across town in half an hour, he'll immediately head for the driveway, an action that seems to be best explained by the hypothesis that he wants to go to the basketball game, believes that his car will get him there, and believes that his car is in the driveway. All of the usual principles of psychological explanation sanction attributing beliefs to Zombie Dave; explaining his action without the attribution of beliefs would be a fearsomely complex task. (It might be

objected that Zombie Dave lacks the external grounding required for belief contents, but we can avoid this problem by stipulating that his environment and history are physically indistinguishable from mine.)

Goldman argues in Section 8 that beliefs, like perceptual states, are typically accompanied by qualia; but much more would be required to conclude that qualia are *essential* to a state's being a belief. (Searle (1990) has given an argument in this direction, but it does not seem to have been widely accepted.) Zombie Dave's beliefs may not be colored by the usual phenomenological tinges, but it seems reasonable to say that they are nevertheless beliefs. Beliefs, unlike qualia, seem to be characterized primarily by the role that they play in the mind's causal economy. (To illustrate the difference, note that it seems *coherent* to be an epiphenomenalist about qualia, whether or not one finds the position plausible; but there seems to be something *conceptually* wrong with the idea that beliefs could be epiphenomenal.) So qualia-free believers like Zombie Dave are quite conceptually coherent, and qualia don't seem to be an essential part of our concept of belief.

Even if one resists the idea that Zombie Dave has beliefs, we can still use him to show that qualia cannot be the primary mechanism in the self-ascription of our mental states. For Zombie Dave ascribes precisely the same mental states to himself as I do! By some process or other, he'll tell you that he thinks that Bob Dylan makes good music. How can this ability for self-ascription be explained? Clearly not by appealing to qualia, for Zombie Dave doesn't have any. The story will presumably have to be told in purely functional terms. But once we have this story in hand, it will apply equally to proud possessors of qualia such as ourselves. The self-ascription mechanisms that Zombie Dave uses are equally the mechanisms that we use; at most, the difference consists in the fact that his ascriptions might be wrong, whereas ours are right. Therefore there is no need to invoke qualia in the explanation of how we ascribe mental states to ourselves. Zombie Dave does the job, presumably, either by reasoning from non-qualitative evidence, or by simply being thrown into the appropriate state. It seems likely that we do it the same way, and that qualia are a red herring.

All this seems to lead to a rather epiphenomenalist view of qualia. Note, for instance, that the argument in the above paragraph doesn't apply only to the self-ascription of beliefs, but also to the self-ascription of qualia; so that qualia don't seem to play a primary role in the process by which we ascribe qualia to ourselves! (Zombie Dave, after all, ascribes himself the same qualia; it's just that he's wrong about it.) I am happy enough with the conclusion that qualia are mostly just along for the ride, but I suspect that Goldman and others will not be. It seems to me that the only way to avoid this conclusion is to deny that Zombie Dave is a conceptual possibility; and the only principled way to deny that Zombie Dave is a conceptual possibility is to allow that functional organization is conceptually constitutive of qualitative content. This is probably a step that Goldman doesn't wish to take, as it would negate many of his conclusions, but there may not be any tenable middle ground between functionalism and epiphenomenalism.

References

Armstrong, D.M. 1968. *A Materialist Theory of the Mind*. Routledge and Kegan Paul.

Dennett, D.C. 1991. *Consciousness Explained*. Little-Brown.

Searle, J.R. 1990. Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences* 13:585-642.

Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:291-315.

Review of *Journal of Consciousness Studies*

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

How does conscious experience emerge from a physical basis? At a first glance, this is *the* question about the mind that most needs answering. So it is curious that those who study the mind professionally have often avoided the question entirely. In psychology, the cognitive revolution did not make consciousness respectable: most cognitive psychologists have stuck to subjects such as learning, memory, and perception instead. Neuroscientists have been known to speculate on the topic, but usually only late at night, after a few drinks. Even philosophers have been curiously diffident. Some have been exercised by the fact that there is a problem, others have been concerned to deny the problem entirely, but the focus of inquiry has remained elsewhere. As in all these fields, serious theories of consciousness have been hard to come by.

But consciousness is making a comeback. In the last few years, popular books on the subject by Francis Crick, Daniel Dennett, and Roger Penrose have provoked widespread discussion. Not everybody agrees on just what the problems are, and almost nobody agrees on where a solution will lie, but almost everyone has an opinion. Philosophers are having a field day, neuroscientists are speculating a little earlier in the evening, and even psychologists have begun to sheepishly utter the dreaded C-word. At conferences on the subject, researchers from different fields are learning to speak each other's language. New books are appearing every month. And there are now three academic journals devoted to the subject.

The latest and flashiest of these journals is the *Journal of Consciousness Studies*. The first issue boasts interviews with such luminaries as Crick and Penrose, a number of conference reports, and refereed articles on consciousness based in areas as diverse as quantum mechanics and Buddhist philosophy. The journal's center of gravity is outside the mainstream, with the editors admitting a preference for "radical" conceptualizations of the subject. The bulk of the first issue is devoted to scientific approaches of one sort or another, but we are promised a discussion of "ethical, spiritual, and social issues" in the future. All in all, the journal is surprisingly accessible to the general reader, and despite its biases it provides a useful look at the state of play in consciousness research *circa* 1994.

One thing that quickly becomes clear is that the most puzzling questions about consciousness are just as puzzling as ever. A distinction is made a number of times between the "easy" problems of consciousness and the "hard" problem. The easy problems are those of finding neural mechanisms and explaining

cognitive functions: the ability to discriminate and categorize environmental stimuli, the capacity to verbally report mental states, the difference between waking and sleeping. The hard problem is that of *experience*: why does all this processing give rise to an experienced inner life at all? While progress is being made on the easy problems, the hard problem remains perplexing. Even Crick and Penrose concede that so far they have little idea how the problem might be solved. They simply hope that if we do enough investigation in neuroscience (for Crick) or in physics (for Penrose), the faint outlines of a solution *might* be revealed. At this point, such remarks are not much more than an expression of faith.

But even without a solution to the hard problem, there is much of interest in the field of consciousness studies. The papers in this issue range from the conventional to the exotic. The most sober paper concerns the "binding problem" in neuroscience - the problem of how the brain integrates different pieces of information about the same object. A popular hypothesis has been that binding is achieved by groups of neurons oscillating in phase, but philosopher Valerie Hardcastle suggests that the evidence here is weak, and recommends a higher-level approach to the issue. In tying together evidence from a number of different fields, this thoughtful paper illustrates the benefits of an interdisciplinary approach to the science of the mind.

It is the quantum-mechanical papers that will draw the most attention. In a hypothesis that has been endorsed by Penrose, anesthesiologist Stuart Hameroff suggests that the key to consciousness lies in microtubules, large molecules found within the cell wall of a neuron. Noting that these are finely poised between the quantum and neural realms - small enough to maintain a state of quantum coherence, but large enough to affect neural activity - Hameroff argues that these could mediate a quantum role in brain function. This is recommended as a solution to the binding problem (through quantum nonlocality), as an explanation of "intuitive" processing (through quantum superposition) and problem-solving (through wavefunction collapse), and even as the key to the problem of free will (through quantum indeterminacy)! Critics will respond that all of these problems can be handled by orthodox methods, but Hameroff's ideas are at least provocative.

Another paper presents experimental evidence that EEG measurements on the brain affect performance on various cognitive tasks. The authors (Nunn, Clarke, and Blott) suggest that the measurement may cause the brain's quantum wave state to collapse! Perhaps this is related to Benjamin Libet's suggestion in a neighbouring paper that the brain has an associated "conscious mental field", which may affect neural function. In both cases, however, the evidence is quite tenuous, and one might ask for a more rigorous theoretical basis for the strong claims that are made. On the other hand, psychologist John Beloff provides a spirited defense of the claims of parapsychology, and argues that the experimental evidence has been dismissed more quickly than it should have been. There is room for a debate here: just when do well-supported prior convictions justify the easy dismissal of evidence for radically different views? No doubt many mainstream scientists will dismiss the evidence of Nunn *et al* and of Beloff without qualms. Is this reasonable?

Two of the most interesting papers relate Eastern thought to the study of the mind. Eleanor Rosch uses some ideas from the Buddhist Madhyamika school in developing a theory of the way that people think about causality, and Robert Forman discusses mystical experiences as revealed in the writings of the

Hindu Upanishads. Rosch brings the Buddhist ideas to bear on issues in cognitive psychology, while Forman uses his study to help adjudicate a debate between "constructivists" and "decontextualists" in religious studies. In papers like these, there is no claim that a theory of consciousness is on offer, but the wide-ranging discussion provides useful insights.

To judge from this issue, one would think that researchers on consciousness come in four varieties: neuroscientists, quantum theorists, parapsychologists, and mystics. This makes for a spicy journal, but it would be nice to see contributions from mainstream cognitive psychology and artificial intelligence in there too. It would be a great pity if too much spice caused conventional researchers to dismiss the journal altogether. But it is pleasant that we are promised discussions in the humanities as well as in science, with papers on cultural aspects of consciousness and on the philosophy of mind slated for future issues. The journal will continue to have something for everybody.

So, the question arises: Has consciousness been explained? Is it likely to be explained any time soon? After a while, I found that I had a curious reaction. The approaches here are too wild for the easy problems, but not wild enough for the hard problem. When it comes to the explanation of cognitive functioning, conventional methods seem to do just fine, and it is hard to see why things like quantum mechanics need to be brought in at all. But when it comes to the hard problem of conscious experience, all the quantum mechanics and parapsychology in the world does not add up to a solution. Why should quantum coherence in microtubules give rise to conscious experience? This question seems no easier to answer than the corresponding question about information-processing in neurons. Radical as the approach is, it is ultimately not radical enough.

Still, a journal like this serves a useful function, both in stimulating thought and in letting a thousand flowers bloom. After all, the problem of conscious experience is so hard that when a solution finally appears, it will probably look crazy. With an unconventional journal on the scene, the required craziness will be given room to flourish.

Absent Qualia, Fading Qualia, Dancing Qualia

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

[[Published in *Conscious Experience*, edited by Thomas Metzinger. Imprint Academic, 1995.]]

1 The principle of organizational invariance

It is widely accepted that conscious experience has a physical basis. That is, the properties of experience (phenomenal properties, or qualia) systematically depend on physical properties according to some lawful relation. There are two key questions about this relation. The first concerns the strength of the laws: are they logically or metaphysically necessary, so that consciousness is nothing "over and above" the underlying physical process, or are they merely contingent laws like the law of gravity? This question about the strength of the psychophysical link is the basis for debates over physicalism and property dualism. The second question concerns the shape of the laws: precisely how do phenomenal properties depend on physical properties? What sort of physical properties enter into the laws' antecedents, for instance; consequently, what sort of physical systems can give rise to conscious experience? It is this second question that I address in this paper.

To put the issue differently, even once it is accepted that experience arises from physical systems, the question remains open: in virtue of what sort of physical properties does conscious experience arise? Some property that brains can possess will presumably be among them, but it is far from clear just what the relevant properties are. Some have suggested biochemical properties; some have suggested quantum-mechanical properties; many have professed uncertainty. A natural suggestion is that when experience arises from a physical system, it does so in virtue of the system's *functional organization*. On this view, the chemical and indeed the quantum substrates of the brain are not directly relevant to the existence of consciousness, although they may be indirectly relevant. What is central is rather the brain's abstract causal organization, an organization that might be realized in many different physical substrates.

In this paper I defend this view. Specifically, I defend a principle of *organizational invariance*, holding that experience is invariant across systems with the same fine-grained functional organization. More precisely, the principle states that given any system that has conscious experiences, then any system that has the same functional organization at a fine enough grain will have qualitatively identical conscious experiences. A full specification of a system's fine-grained functional organization will fully determine any conscious experiences that arise.

To clarify this, we must first clarify the notion of functional organization. This is best understood as the *abstract pattern of causal interaction* between the components of a system, and perhaps between these components and external inputs and outputs. A functional organization is determined by specifying (1) a number of abstract components, (2) for each component, a number of different possible states, and (3) a system of dependency relations, specifying how the states of each component depends on the previous states of all components and on inputs to the system, and how outputs from the system depend on previous component states. Beyond specifying their number and their dependency relations, the nature of the components and the states is left unspecified.

A physical system *realizes* a given functional organization when the system can be divided into an appropriate number of physical components each with the appropriate number of possible states, such that the causal dependency relations between the components of the system, inputs, and outputs precisely reflect the dependency relations given in the specification of the functional organization. A given functional organization can be realized by diverse physical systems. For example, the organization realized by the brain at the neural level might in principle be realized by a silicon system.

A physical system has functional organization at many different levels, depending on how finely we individuate its parts and on how finely we divide the states of those parts. At a coarse level, for instance, it is likely that the two hemispheres of the brain can be seen as realizing a simple two-component organization, if we choose appropriate interdependent states of the hemispheres. It is generally more useful to view cognitive systems at a finer level, however. For our purposes I will always focus on a level of organization fine enough to determine the behavioral capacities and dispositions of a cognitive system. This is the role of the "fine enough grain" clause in the statement of the organizational invariance principle; the level of organization relevant to the application of the principle is one fine enough to determine a system's behavioral dispositions. In the brain, it is likely that the neural level suffices, although a coarser level might also work. For the purposes of illustration I will generally focus on the neural level of organization of the brain, but the arguments generalize.

Strictly speaking, for the purposes of the invariance principle we must require that for two systems to share their functional organization, they must be in corresponding states at the time in question; if not for this requirement, my sleeping twin might count as sharing my organization, but he certainly does not share my experiences. When two systems share their organization at a fine enough grain (including the requirement that they be in corresponding states), I will say that they are *functionally isomorphic* systems, or that they are *functional isomorphs*. The invariance principle holds that any functional isomorph of a conscious system has experiences that are qualitatively identical to those of the original system.

2 Absent qualia and inverted qualia

The principle of organizational invariance is far from universally accepted. Some have thought it likely that for a system to be conscious it must have the right sort of biochemical makeup; if so, a metallic robot

or a silicon-based computer could never have experiences, no matter what its causal organization. Others have conceded that a robot or a computer might be conscious if it were organized appropriately, but have held that it might nevertheless have experiences quite different from the kind that we have. These two sorts of objections are often known as the *absent qualia* and *inverted qualia* objections to broadly functionalist theories of consciousness.

Arguments for the absent qualia objection usually consist in the description of a system that realizes whatever functional organization might be specified, but that is so outlandish that it is natural to suppose that it lacks conscious experience. For example, Block (1980) points out that the functional organization of the brain might be instantiated by the population of China, if they were organized appropriately, and argues that it is bizarre to suppose that this would somehow give rise to a group mind. In a similar way, John Searle (1980) notes that a given organization might be realized by "a sequence of water-pipes, or a set of wind-machines", but argues that these systems would not be conscious.

Arguments for the inverted qualia objection are often illustrated by considerations about experiences of color. According to this line of argument (taken by Shoemaker (1982) and Horgan (1984), among others), it is possible that a system might make precisely the same color discriminations that I do, but that when confronted by red objects it has the kind of experience that I have when confronted by blue objects. Further, it is argued that this might happen even when the systems are functionally isomorphic. If this argument succeeds, then even if the appropriate functional organization suffices for the existence of conscious experiences, it does not determine their specific nature. Instead, the specific nature of experiences must be dependent on non-organizational properties, such as specific neurophysiological properties.

Sometimes these arguments are intended as arguments for "possibility" only in some weak sense, such as logical or metaphysical possibility. These less ambitious forms of the arguments are the most likely to be successful. It seems difficult to deny that the absent qualia and inverted qualia scenarios are at least intelligible. With the aid of certain assumptions about possibility, this intelligibility can be extended into an argument for the logical and perhaps the metaphysical possibility of the scenarios. If successful, even these less ambitious arguments would suffice to refute some strong versions of functionalism, such as analytic functionalism and the view that phenomenal properties are identical to functional properties.

In the present paper I am not concerned with the logical or metaphysical possibility of these scenarios, however, but rather with their *empirical* (or natural, or nomological) possibility. The mere logical or metaphysical possibility of absent qualia is compatible with the claim that in the actual world, whenever the appropriate functional organization is realized, conscious experience is present. By analogy: many have judged it logically possible that a *physical* replica of a conscious system might lack conscious experience, while not wishing to deny that in the actual world, any such replica will be conscious. It is the claim about empirical possibility that is relevant to settling the issue at hand, which concerns a possible lawful relation between organization and experience. Mere intelligibility does not bear on this, any more than the intelligibility of a world without relativity can falsify Einstein's theory.

On the question of empirical possibility, the success of the absent qualia and inverted qualia arguments is unclear. To be sure, many have found it *counterintuitive* that the population of China might give rise to conscious experience if organized appropriately. The natural reply, however, is that it seems equally counterintuitive that a mass of 10^{11} appropriately organized neurons should give rise to consciousness, and yet it happens. Intuition is unreliable as a guide to empirical possibility, especially where a phenomenon as perplexing as conscious experience is concerned. If a brain can do the job of enabling conscious experience, it is far from obvious why an appropriately organized population, or indeed an appropriate organized set of water-pipes, could not.

The debate over absent and inverted qualia tends to produce a stand-off, then. Both proponents and opponents claim intuitions in support of their positions, but there are few grounds on which to settle the debate between them. Both positions seem to be epistemic possibilities, and due to the notorious difficulties in collecting experimental evidence about conscious experience, things might seem likely to stay that way.

I believe that the stand-off can be broken, and in this paper I will offer considerations that offer strong support to the principle of organizational invariance, suggesting that absent qualia and inverted qualia are empirically impossible. These arguments involve thought-experiments about gradual neural replacement, and take the form of a *reductio*. The first thought-experiment demonstrates that if absent qualia are possible, then a phenomenon involving what I will call *Fading Qualia* is possible; but I will argue that we have good reason to believe that Fading Qualia are impossible. The second argument has broader scope and is more powerful, demonstrating that if absent qualia *or* inverted qualia are possible, then a phenomenon involving what I will call *Dancing Qualia* is possible; but I will argue that we have even better reason to believe that Dancing Qualia are impossible. If the arguments succeed, we have good reason to believe that absent and inverted qualia are impossible, and that the principle of organizational invariance is true.

These arguments do not constitute conclusive *proof* of the principle of organizational invariance. Such proof is generally not available in the domain of conscious experience, where for familiar reasons one cannot even disprove the hypothesis that there is only one conscious being. But even in the absence of proof, we can bring to bear arguments for the plausibility and implausibility of different possibilities, and not all possibilities end up equal. I use these thought-experiments as a *plausibility argument* for the principle of organizational invariance, by showing that the alternatives have implausible consequences. If an opponent wishes to hold on to the possibility of absent or inverted qualia she can still do so, but the thought-experiments show that the cost is higher than one might have expected.

Perhaps it is useful to see these thought-experiments as playing a role analogous to that played by the "Schrödinger's cat" thought-experiment in the interpretation of quantum mechanics. Schrödinger's thought-experiment does not deliver a decisive verdict in favor of one interpretation or another, but it brings out various plausibilities and implausibilities in the interpretations, and it is something that every interpretation must ultimately come to grips with. In a similar, any theory of consciousness must ultimately come to grips with the Fading and Dancing Qualia scenarios, and some will handle them better than others. In this way, the virtues and drawbacks of various theories are clarified.

3 Fading Qualia

The first scenario that I will present is relatively familiar,[*] but it is important to analyze it correctly, and it is a necessary preliminary to the more powerful second argument. In this thought-experiment, we assume for the purposes of *reductio* that absent qualia are empirically possible. It follows that there can be a system with the same functional organization as a conscious system (such as me), but which lacks conscious experience entirely due to some difference in non-organizational properties. Without loss of generality, suppose that this is because the system is made of silicon chips rather than neurons. Call this functional isomorph Robot. The causal patterns in Robot's processing system are the same as mine, but there is nothing it is like to be Robot.

*[[Neural replacement scenarios along the lines discussed in this section are discussed by Pylyshyn (1980), Savitt (1980), Cuda (1985), and Searle (1992), among others.]]

Given this scenario, we can construct a series of cases intermediate between me and Robot such that there is only a very small change at each step and such that functional organization is preserved throughout. We can imagine, for instance, replacing a certain number of my neurons by silicon chips. In the first such case, only a single neuron is replaced. Its replacement is a silicon chip that performs precisely the same local function as the neuron. We can imagine that it is equipped with tiny transducers that take in electrical signals and chemical ions and transforms these into a digital signal upon which the chip computes, with the result converted into the appropriate electrical and chemical outputs. As long as the chip has the right input/output function, the replacement will make no difference to the functional organization of the system.

In the second case, we replace two neighboring neurons with silicon chips. This is just as in the previous case, but once both neurons are replaced we can eliminate the intermediary, dispensing with the awkward transducers and effectors that mediate the connection between the chips and replacing it with a standard digital connection. Later cases proceed in a similar fashion, with larger and larger groups of neighboring neurons replaced by silicon chips. Within these groups, biochemical mechanisms have been dispensed with entirely, except at the periphery. In the final case, every neuron in the system has been replaced by a chip, and there are no biochemical mechanisms playing an essential role. (I abstract away here from detailed issues concerning whether, for instance, glial cells play a non-trivial role; if they do, they will be components of the appropriate functional organization, and will be replaced also.)

We can imagine that throughout, the internal system is connected to a body, is sensitive to bodily inputs, and produces motor movements in an appropriate way, via transducers and effectors. Each system in the sequence will be functionally isomorphic to me at a fine enough grain to share my behavioral dispositions. But while the system at one end of the spectrum is me, the system at the other end is essentially a copy of Robot.

To fix imagery, imagine that as the first system I am having rich conscious experiences. Perhaps I am at

a basketball game, surrounded by shouting fans, with all sorts of brightly-colored clothes in my environment, smelling the delicious aroma of junk food and perhaps suffering from a throbbing headache. Let us focus in particular on the bright red and yellow experiences I have when I watch the players' uniforms. ("Red experience" should be taken as shorthand for "color experience of the kind I usually have when presented with red objects", and so on throughout.) The final system, Robot, is in the same situation, processing the same inputs and producing similar behavior, but by hypothesis is experiencing nothing at all.

The question arises: *What is it like to be the systems in between?* For those systems intermediate between me and Robot, what, if anything, are they experiencing? As we move along the spectrum of cases, how does conscious experience vary? Presumably the very early cases have experiences much like mine, and the very late cases have little or no experience, but what of the cases in the middle?

Given that Robot, at the far end of the spectrum, is not conscious, it seems that one of two things must happen along the way. Either consciousness gradually fades over the series of cases, before eventually disappearing, or somewhere along the way consciousness suddenly blinks out, although the preceding case had rich conscious experiences. Call the first possibility *Fading Qualia* and the second *Suddenly Disappearing Qualia*.

On the second hypothesis, the replacement of a single neuron could be responsible for the vanishing of an entire field of conscious experience. If so, we could switch back and forth between a neuron and its silicon replacement, with a field of experience blinking in and out of existence on demand. This seems antecedently implausible, if not entirely bizarre. If Suddenly Disappearing Qualia were possible, there would be brute discontinuities in the laws of nature unlike those we find anywhere else.[*] Any specific point for qualia to suddenly disappear (50 percent neural? 25 percent?) would be quite arbitrary. We might even run the experiment at a finer grain within the neuron, so that ultimately the replacement of a few molecules produces a sudden disappearance of experience. As always in these matters, the hypothesis cannot be disproved, but its antecedent plausibility is very low.

*[[One might argue that there are situations in nonlinear dynamics in which one magnitude depends sensitively on another, with large changes in the first arising from small changes in the second. But in these cases the dependence is nevertheless continuous, so there will be intermediate cases in which the dependent magnitude takes on intermediate values; the analogy therefore leads to Fading Qualia, below. And in any case, the sensitive dependence in these cases generally arise from the compound effects of a number of more basic gradual dependencies. In all fundamental laws known to date, the dependence of one magnitude on another is continuous in this fashion, and there is no way to compound continuity into discontinuity. Suddenly Disappearing Qualia, in contrast to nonlinear dynamics, would therefore require brute discontinuities in fundamental laws.]]

This leaves the first hypothesis, Fading Qualia. To get a fix on this hypothesis, consider a system halfway along the spectrum between me and Robot, after consciousness has degraded considerably but before it has gone altogether. Call this system Joe. What is it like to be Joe? Joe, of course, is functionally isomorphic to me. He says all the same things about his experiences as I do about mine. At the basketball game, he exclaims about the vivid bright red and yellow uniforms of the basketball players.

By hypothesis, though, Joe is not having bright red and yellow experiences at all. Instead, perhaps he is experiencing tepid pink and murky brown. Perhaps he is having the faintest of red and yellow experiences. Perhaps his experiences have darkened almost to black. There are various conceivable ways in which red experiences might gradually transmute to no experience, and probably more ways that we cannot conceive. But presumably in each of these transmutation scenarios, experiences stop being *bright* before they vanish (otherwise we are left with the problem of Suddenly Disappearing Qualia). Similarly, there is presumably a point at which subtle distinctions in my experience are no longer present in an intermediate system's experience; if we are to suppose that all the distinctions in my experience are present right up until a moment when they simultaneously vanish, we are left with another version of Suddenly Disappearing Qualia.

For specificity, then, let us imagine that Joe experiences faded pink where I see bright red, with many distinctions between shades of my experience no longer present in shades of his experience. Where I am having loud noise experiences, perhaps Joe is experiencing only a distant rumble. Not everything is so bad for Joe: where I have a throbbing headache, he only has the mildest twinge.

The crucial point here is that Joe is systematically *wrong* about everything that he is experiencing. He certainly *says* that he is having bright red and yellow experiences, but he is merely experiencing tepid pink. If you ask him, he will claim to be experiencing all sorts of subtly different shades of red, but in fact many of these are quite homogeneous in his experience. He may even complain about the noise, when his auditory experience is really very mild. Worse, on a functional construal of judgment, Joe will even *judge* that he has all these complex experiences that he in fact lacks. In short, Joe is utterly out of touch with his conscious experience, and is incapable of getting in touch.

There is a significant implausibility here. This is a being whose rational processes are functioning and who is in fact *conscious*, but who is completely wrong about his own conscious experiences. Perhaps in the extreme case, when all is dark inside, it is reasonable to suppose that a system could be so misguided in its claims and judgments - after all, in a sense there is nobody in there to be wrong. But in the intermediate case, this is much less plausible. In every case with which we are familiar, conscious beings are generally capable of forming accurate judgments about their experience, in the absence of distraction and irrationality. For a sentient, rational being that is suffering from no functional pathology to be so systematically out of touch with its experiences would imply a strong dissociation between consciousness and cognition. We have little reason to believe that consciousness is such an ill-behaved phenomenon, and good reason to believe otherwise.

To be sure, Fading Qualia may be *logically* possible. Arguably, there is no contradiction in the notion of a system that is so wrong about its experiences. But logical possibility and empirical possibility are different things. One of the most salient empirical facts about conscious experience is that when a conscious being with the appropriate conceptual sophistication has experiences, it is at least capable of forming reasonable judgments about those experiences. Perhaps there are some cases where judgment is impaired due to a malfunction in rational processes, but this is not such a case. Joe's processes are *functioning* as well as mine -- by hypothesis, he is functionally isomorphic. It is just that he happens to be

completely misguided about his experience.

There are various cases of fading qualia in everyday life, of course. Think of what happens when one is dropping off to sleep; or think of moving back along the evolutionary chain from people to trilobites. In each case, as we move along a spectrum of cases, conscious experience gradually fades away. But in each of these cases, the fading is accompanied by a corresponding change in *functioning*. When I become drowsy, I do not believe that I am wide awake and having intense experiences (unless perhaps I start to dream, in which case I very likely *am* having intense experiences). The lack of richness in a dog's experience of color accompanies a corresponding lack of discriminatory power in a dog's visual mechanisms. These cases are quite unlike the case under consideration, in which experience fades while functioning stays constant. Joe's mechanisms can still discriminate subtly different wavelengths of light, and he certainly judges that such discriminations are reflected in his experience, but we are to believe that his experience does not reflect these discriminations at all.

Searle[*] discusses a thought-experiment like this one, and suggests the following possibility:

...as the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when the doctors test your vision, you hear them say, "We are holding up a red object in front of you; please tell us what you see." You want to cry out, "I can't see anything. I'm going totally blind." But you hear your voice saying in a way that is completely out of your control, "I see a red object in front of me." If we carry the thought-experiment out to the limit, we get a much more depressing result than last time. We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same.

*[[Searle, *The Rediscovery of the Mind*, pp. 66-67.]]

Here, Searle embraces the possibility of Fading Qualia, but suggests that such a system need not be systematically mistaken in its beliefs about its experience. The system might have true beliefs about its experience, but beliefs that are impotent to affect its behavior.[*]

*[[Searle also raises the possibility that upon silicon replacement, the system might be slowly reduced to paralysis, or have its functioning otherwise impaired. Such a scenario is irrelevant to the truth of the invariance principle, however, which applies only to systems with the appropriate functional organization. If a silicon system does not duplicate the organization of the original system, the principle does not even come into play.]]

It seems that this possibility can be ruled out, however. There is simply no room in the system for any new beliefs to be formed. Unless one is a dualist of a very strong variety, beliefs must be reflected in the functioning of a system - *perhaps* not in behavior, but at least in some process. But this system is identical to the original system (me) at a fine grain. There is no room for new beliefs like "I can't see

anything", new desires like the desire to cry out, and other new cognitive states such as amazement. Nothing in the physical system can correspond to that amazement. There is no room for it in the neurons, which after all are identical to a subset of the neurons supporting the usual beliefs; and Searle is surely not suggesting that the silicon replacement is itself supporting the new beliefs! Failing a remarkable, magical interaction effect between neurons and silicon - and one that does not manifest itself anywhere in processing, as organization is preserved throughout - such new beliefs will not arise.

While it might just seem plausible that an organization-preserving change from neurons to silicon might twist a few experiences from red to blue, a change in beliefs from "Nice basketball game" to "I seem to be stuck in a bad horror movie!" is of a different order of magnitude. If such a major change in cognitive contents were not mirrored in a change in functional organization, cognition would float free of internal functioning like a disembodied mind. If the contents of cognitive states supervened on physical states at all, they could do so only by the most arbitrary and capricious of rules (if this organization in neurons, then "pretty colors!"; if this organization in silicon, then "Alas!").

It follows that the possibility of Fading Qualia requires either a bizarre relationship between belief contents and physical states, or the possibility of beings that are massively mistaken about their own conscious experiences despite being fully rational. Both of these hypotheses are significantly less plausible than the hypothesis that rational conscious beings are generally correct in their judgments about their experiences. A much more reasonable hypothesis is therefore that when neurons are replaced, qualia do not fade at all. A system like Joe, in practice, will have conscious experiences just as rich as mine. If so, then our original assumption was wrong, and the original isomorph, Robot, has conscious experiences.

This thought-experiment can be straightforwardly extended to other sorts of functional isomorphs, including those that differ in shape, size, and physical makeup. All we need do is construct a sequence of intermediate cases, each with the same functional organization. In each case the conclusion is the same. If such a system is not conscious, then there exists an intermediate system that is conscious, has faded experiences, and is completely wrong about its experiences. Unless we are prepared to accept this massive dissociation between consciousness and cognition, the original system must have been conscious after all.

We can even extend the reasoning straightforwardly to the case of an appropriately-organized population: we simply need to imagine neurons replaced one-by-one with tiny homunculi, ending up with a network of homunculi that is essentially equivalent to the population controlling a robot. (If one objects to tiny homunculi, they can be external and of normal size, as long as they are equipped with appropriate radio connections to the body when necessary.) Precisely the same considerations about intermediate cases arise. One can also imagine going from a multiple-homunculi case to a single-homunculus case, yielding something like Searle's "Chinese room" example. We need only suppose that the homunculi gradually "double up" on their tasks, leaving written records of the state of each component, until only a single homunculus does all the work. If the causal organization of the original system is preserved, even if it is only among a system of marks on paper, then the same arguments suggest that the system will have experiences. (Of course, we should not expect the homunculus itself to

have the experiences; it is merely acting as a sort of causal facilitator.)

If Absent Qualia are possible, then Fading Qualia are possible. But I have argued above that it is very unlikely that Fading Qualia are possible. It follows that it is very unlikely that absent qualia are possible.

Some might object that these thought-experiments are the stuff of science fiction rather than the stuff of reality, and point out that this sort of neural replacement would be quite impossible in practice. But although it might be technologically impossible, there is no reason to believe that the neural replacement scenario should be nomologically impossible. We already have prosthetic arms and legs. Prosthetic eyes lie within the foreseeable future, and there seems to be no reason why a prosthetic neuron is impossible in principle. Even if it were impossible for some technical reason (perhaps there would not be enough room for a silicon replacement to do its work?), it is entirely unclear what bearing this technical fact would have on the principled force of the thought-experiment. There will surely be *some* systems between which gradual replacement is possible; will the objector hold that the invariance principle holds for those systems, but no other? If so, the situation seems quite arbitrary; if not, then there must be a deeper objection available.

Others might object that no silicon replacement could perform even the local function of a neuron, perhaps because neural function is uncomputable. There is little evidence for this, but it should be noted that even if it is true, it does not affect the argument for the invariance principle. If silicon really could not even duplicate the *function* of a neural system, then a functional isomorph made of silicon would be impossible, and the assessment of silicon systems would simply be irrelevant to the invariance principle. To evaluate the truth of the principle, it is only functionally isomorphic systems that are relevant.

Another objection notes that there are actual cases in which subjects are seriously mistaken about their experiences. For example, in cases of blindness denial, subjects believe that they are having visual experiences when they likely have none. In these cases, however, we are no longer dealing with fully rational systems. In systems whose belief-formation mechanisms are impaired, anything goes. Such systems might believe that they are Napoleon, or that the moon is pink. My "faded" isomorph Joe, by contrast, is a fully rational system, whose cognitive mechanisms are functioning just as well as mine. In conversation, he seems perfectly sensible. We cannot point to any unusually poor inferential connections between his beliefs, or any systematic psychiatric disorder that is leading his thought processes to be biased toward faulty reasoning. Joe is an eminently thoughtful, reasonable person, who exhibits none of the confabulatory symptoms of those with blindness denial. The cases are therefore disanalogous. The plausible claim is not that no system can be massively mistaken about its experiences, but that no rational system whose cognitive mechanisms are unimpaired can be so mistaken. Joe is certainly a rational system whose mechanisms are working as well as mine, so the argument is unaffected.

Some object that this argument has the form of a Sorites or "slippery-slope" argument, and observe that these arguments are notoriously suspect. Using a Sorites argument, we can "show" that even a grain of sand is a heap; after all, a million grains of sand form a heap, and if we take a single grain away from a heap we still have a heap. This objection is based on a superficial reading of the thought-experiment,

however. Sorites arguments gain their force by ignoring the fact that some apparent dichotomy is in fact a continuum; there are all sorts of vague cases between heaps and non-heaps, for instance. The Fading Qualia argument, by contrast, explicitly accepts the possibility of a continuum, but argues that intermediate cases are impossible for independent reasons. The argument is therefore not a Sorites argument.

Ultimately, the only tenable way for an opponent of organizational invariance to respond to this argument is to bite the bullet and accept the possibility of Fading Qualia, and the consequent possibility that a rational conscious system might be massively mistaken about its experience, or perhaps to bite another bullet and accept Suddenly Disappearing Qualia and the associated brute discontinuities. These positions seem much less plausible than the alternative, other things being equal, but they are the only way to avoid it. But there is worse to come: the argument to follow provides an even more powerful case against the possibility of absent qualia, so opponents of organizational invariance cannot rest easily.

4 Dancing Qualia

If the Fading Qualia argument succeeds, it establishes that functional isomorphs of a conscious system will have conscious experience, but it does not establish that isomorphs have the *same* sort of conscious experience. The preceding argument has no bearing on the possibility of inverted qualia. For all that has gone before, where I am having a red experience, my silicon functional isomorph might be having a blue experience, or some other kind of experience that is quite foreign to me.

One might think that the Fading Qualia argument could be directly adapted to provide an argument against the possibility of inverted qualia, but that strategy fails. If I have a red experience and my functional isomorph has a blue experience, there is no immediate problem with the idea of intermediate cases with intermediate experiences. These systems might be simply suffering from milder cases of qualia inversion, and are no more problematic than the extreme case. These systems will not be systematically wrong about their experiences. Where they claim to experience distinctions, they may really be experiencing distinctions; where they claim to be having intense experiences, they may still be having intense experiences. To be sure, the experiences they call "red" differ from those I call "red", but this is already an accepted feature of the usual inversion case. The difference between these cases and the Fading Qualia cases is that these cases preserve the *structure* of experience throughout, so that their existence implies no implausible dissociation between experience and cognition.

Nevertheless, a good argument against the possibility of inverted qualia can be found in the vicinity. Once again, for the purposes of *reductio*, assume that inverted qualia are empirically possible. Then there can be two functionally isomorphic systems that are having different experiences. Suppose for the sake of illustration that these systems are me, having a red experience, and my silicon isomorph, having a blue experience (there is a small caveat about generality, which I discuss below).

As before, we construct a series of cases intermediate between me and my isomorph. Here, the argument takes a different turn. We need not worry about the *way* in which experiences change as we move along

the series. All that matters is that there must be two points *A* and *B* in this series, such that no more than one-tenth of the system is replaced between *A* and *B*, and such that *A* and *B* have significantly different experiences. To see that this must be the case, we need only consider the points at which 10 percent, 20 percent, and so on up to 90 percent of the brain has been replaced. Red and blue are sufficiently different experiences that some neighboring pairs here *must* be significantly different (that is, different enough that the difference would be noticeable if they were experienced by the same person); there is no way to get from red to blue by ten non-noticeable jumps.

There must therefore be two systems that differ in at most one-tenth of their internal makeup, but that have significantly different experiences. For the purposes of illustration, let these systems be me and Bill. Where I have a red experience, Bill has a slightly different experience. We may as well suppose that Bill sees blue; perhaps his experience will be more similar to mine than that, but that makes no difference to the argument. The two systems also differ in that where there are neurons in some small region of my brain, there are silicon chips in Bill's brain. This substitution of a silicon circuit for a neural circuit is the only physical difference between me and Bill.

The crucial step in the thought-experiment is to take a silicon circuit just like Bill's and install it in my head as a *backup circuit*. This circuit will be functionally isomorphic to a circuit already present in my head. We equip the circuit with transducers and effectors so that it can interact with the rest of my brain, but we do not hook it up directly. Instead, we install a *switch* that can switch directly between the neural and silicon circuits. Upon flipping the switch, the neural circuit becomes irrelevant and the silicon circuit takes over. We can imagine that the switch controls the points of interface where the relevant circuits affects the rest of the brain. When it is switched, the connections from the neural circuit are pushed out of the way, and the silicon circuit's effectors are attached. (We might imagine that the transducers for both circuits are attached the entire time, so that the state of both circuits evolves appropriately, but so that only one circuit at a time plays a role in processing. We could also run a similar experiment where both transducers and effectors are disconnected, to ensure that the backup circuit is entirely isolated from the rest of the system. This would change a few details, but the moral would be the same.)

Immediately after flipping the switch, processing that was once performed by the neural circuit is now performed by the silicon circuit. The flow of control within the system has been redirected. However, my functional organization is exactly the same as it would have been if we had not flipped the switch. The only relevant difference between the two cases is the physical makeup of one circuit within the system. There is also a difference in the physical makeup of another "dangling" circuit, but this is irrelevant to functional organization, as it plays no role in affecting other components of the system and directing behavior.

What happens to my experience when we flip the switch? Before installing the circuit, I was experiencing red. After we install it but before we flip the switch, I will presumably still be experiencing red, as the only difference is the addition of a circuit that is not involved in processing in any way; for all the relevance it has to my processing, I might as well have eaten it. *After* flipping the switch, however, I am more or less the same system as Bill. The only difference between Bill and me now is that I have a causally irrelevant neural circuit dangling from the system (we might even imagine that the circuit is

destroyed when the switch is flipped). Bill, by hypothesis, was enjoying a blue experience. After the switch, then, I will have a blue experience too.

What will happen, then, is that my experience will change "before my eyes". Where I was once experiencing red, I will now experience blue. All of a sudden, I will have a *blue* experience of the apple on my desk. We can even imagine flipping the switch back and forth a number of times, so that the red and blue experiences "dance" before my eyes.

This might seem reasonable at first - it is a strangely appealing image - but something very odd is going on here. My experiences are switching from red to blue, but *I do not notice any change*. Even as we flip the switch a number of times and my qualia dance back and forth, I will simply go about my business, not noticing anything unusual. My functional organization remains normal throughout. In particular, my functional organization after flipping the switch evolves just as it would have if the switch had not been flipped. There is no special difference in my behavioral dispositions. I am not suddenly disposed to say "Hmm! Something strange is going on!". There is no room for a sudden start, for an exclamation, or even for a distraction of attention. My cognitive organization is just as it usually is, and in particular is precisely as it would have been had the switch not been flipped.

Certainly, on any functional construal of judgment, it is clear that I do not make any novel judgments due to the flip. Even if one were to dispute a functional account of judgment, it is extremely implausible that a simple organization-preserving replacement of a neural circuit by a silicon circuit could be responsible for the addition of significant new judgments such as "My qualia just flipped". As in the case of Fading Qualia, there is simply no room for such a change to take place, unless it is in an accompanying Cartesian disembodied mind.

We are therefore led once more into a *reductio ad absurdum*. It seems entirely implausible to suppose that my experiences could change in such a significant way, even with me paying full attention, without my being able to notice the change. It would suggest once again an extreme dissociation between consciousness and cognition. If this kind of thing could happen, then psychology and phenomenology would be radically out of step, much further out of step than even the Fading Qualia scenario would imply.

This "Dancing Qualia" scenario may be logically possible (although the case is so extreme that it seems *only just* logically possible), but that does not mean we should take it seriously as an empirical possibility, any more than we should take seriously the possibility that the world was created five minutes ago. As an empirical hypothesis, it is far more plausible that when one's experiences change significantly, then as long as one is rational and paying attention, one should be able to notice the change. If not, then consciousness and cognition are tied together by only the most slender of threads.

Indeed, if we are to suppose that Dancing Qualia are empirically possible, we are led to a worrying thought: they might be *actual*, and happening to us all the time. The physiological properties of our functional mechanisms are constantly changing. The functional properties of the mechanisms are

reasonably robust; one would expect that this robustness would be ensured by evolution. But there is no adaptive reason for the non-functional properties to stay constant. From moment to moment there will certainly be changes in low-level molecular properties. Properties such as position, atomic makeup, and so on can change while functional role is preserved, and such change is almost certainly going on constantly.

If we allow that qualia are dependent not just on functional organization but on implementational details, it may well be that *our* qualia are in fact dancing before our eyes all the time. There seems to be no principled reason why a change from neurons to silicon should make a difference while a change in neural realization should not; the only place to draw a *principled* line is at the functional level. The reason why we doubt that such dancing is taking place in our own cases is that we accept the following principle: when one's experiences change significantly, one can notice the change. If we were to accept the possibility of Dancing Qualia in the original case, we would be discarding this principle, and it would no longer be available as a defense against skepticism even in the more usual cases.

It is not out of the question that we could actually perform such an experiment. Of course the practical difficulties would be immense, but at least in principle, one could install such a circuit in me and *I* could see what happened, and report it to the world. But of course there is no point performing the experiment: we know what the result will be. I will report that my experience stayed the same throughout, a constant shade of red, and that I noticed nothing untoward. I will become even more convinced than I was before that qualia are determined by functional organization. Of course this will not be a *proof*, but the evidence will be hard to seriously dispute.

I conclude that by far the most plausible hypothesis is that replacement of neurons while preserving functional organization will preserve qualia, and that experience is wholly determined by functional organization.

There are a few small loopholes that the argument leaves open, but none of the loopholes leads to an attractive position. For example, while the Dancing Qualia scenario is straightforwardly extendible to most sorts of functional isomorphs, there are a couple of exceptions involving speed and history. If an isomorph is much faster or slower than the original system, we cannot simply substitute a circuit from one system into the other and expect everything to function normally. We can still perform the experiment on a slowed-down or speeded-up version of the system in question, however, so at most we have left open the possibility that a change in speed might invert qualia. A similar loophole is left open for physical isomorphs that differ in their *history*: perhaps if I was born in the Southern Hemisphere I experience green, whereas a physical twin born in the North would experience red. History cannot be varied in a Dancing Qualia scenario (although it can be varied in a Fading Qualia scenario), so the argument does not bear on the hypothesis that qualia supervene on the past.

But neither of these hypotheses were very plausible in the first place. It is reasonable that history should affect our qualia by affecting our physical structure, but the history-dependence required above would be much stronger: there would in effect be a "nonlocal" effect of distal history on present qualia, unmediated

by anything in physical structure or nearby in space and time. As for speed, it would seem quite arbitrary that a change inspeed would invert qualia when nothing else could. The hypotheses here are coherent, but there is little reason to embrace them.

Another small caveat is that the argument does not refute the possibility of a very mild spectrum inversion. Between dark red and a slightly darker red, for instance, there may be nine intermediate shades such that no two neighboring shades are distinguishable. In such a case the Dancing Qualia scenario is not a problem; if the system notices no difference on flipping the switch, that is just what we would expect.

Of course, there is nothing special about the figure of one-tenth as the amount of difference between two neighboring systems. But we cannot make the figure too high. If we made it as high as one half, we would run into problems with personal identity: it might reasonably be suggested that upon flipping the switch, we are creating a new person, and it would not be a problem that the new person noticed no change. Perhaps we might go as high as 20 percent or 25 percent without such problems; but that would still allow the possibility of very mild inversions, the kind that could be composed of four or five unnoticeable changes. We can reduce the impact of this worry, however, by noting that it is very unlikely that experience depends equally on all areas of the brain. If color experience depends largely on a small area of the visual cortex, say, then we could perform any qualia inversion in one fell swoop while only replacing a small portion of the system, and the argument would succeed against even the mildest noticeable qualia inversion.

In any case, the possibility of a mild underdetermination of experience by organization is an unthreatening one. If we wished, we could accept it, noting that any differences between isomorphs would be so slight as to be uninteresting. More likely, we can note that this would seem an odd and unlikely way for the world to be. It would seem reasonable that experiences should be invertible across the board, or not invertible at all, but why should the world be such that a small inversion is possible but nothing more? This would seem quite arbitrary. We cannot rule it out, but it is not a hypothesis with much antecedent plausibility.

In a similar way, the argument leaves open the loophole that *unattended* qualia might be invertible. If we are not attending to the fringes of our visual field, for example, a qualia inversion might take place there without our noticing. But to exploit this loophole would leave one in the unattractive position that qualia are organizationally invariant when they are central enough in one's attention, but dependent on other features when they are not. (Presumably an inverted green experience on the fringe will flip back to red when one attends to it?) Such an asymmetric position would be theoretically unsatisfying in the extreme.

It should be noted that the Dancing Qualia argument works just as well against the possibility of absent qualia as against that of inverted qualia. If absent qualia are possible, then on the path to absent qualia we can find two slightly different systems whose experience differs significantly, and we can install a backup circuit in the same way. As before, the hypothesis implies that switching will cause my qualia to dance before my eyes, from vivid to tepid and back, without my ever noticing any change. This is

implausible for the same reasons as before, so we have good reason to believe that absent qualia are impossible.

Overall, the Dancing Qualia argument seems to make an even more convincing case against absent qualia than the Fading Qualia argument does, although both have a role to play. Where an opponent might bite the bullet and accept the possibility of Fading Qualia, Dancing Qualia are an order of magnitude more difficult to accept. The very immediacy of the switch makes a significant difference, as does the fact that the subject cannot notice something so striking and dynamic. The possibility of Fading Qualia would imply that some systems are out of touch with their conscious experience, but Dancing Qualia would establish a much stranger gap.

5 Nonreductive functionalism

To summarize: we have established that if absent qualia are possible, then Fading Qualia are possible; if inverted qualia are possible, then Dancing Qualia are possible; and if absent qualia are possible, then Dancing Qualia are possible. But it is implausible that Fading Qualia are possible, and it is extremely implausible that Dancing Qualia are possible. It is therefore extremely implausible that absent qualia and inverted qualia are possible. It follows that we have good reason to believe that the principle of organizational invariance is true, and that functional organization fully determines conscious experience.

It should be noted that these arguments do not establish functionalism in the strongest sense, as they establish at best that absent and inverted qualia are empirically (naturally, nomologically) impossible. There are two reasons why the arguments cannot be extended into an argument for logical or metaphysical impossibility. First, both Fading Qualia and Dancing Qualia seem to be intelligible hypotheses, even if they are very implausible. Some might dispute their logical possibility, perhaps holding that it is constitutive of qualia that subjects can notice differences between them. This conceptual intuition would be controversial, but in any case, even if we were to accept the logical impossibility of Fading and Dancing Qualia, there is a second reason why these arguments do not do not establish the logical or metaphysical determination of conscious experience by functional organization.

To see this second reason, note that the arguments take as an *empirical* premise certain facts about the distribution of functional organization in physical systems: that I have conscious experiences of a certain kind, or that some biological systems do. If we established the logical impossibility of Fading and Dancing Qualia, this might establish the logical necessity of the *conditional*: if one system with fine-grained functional organization *F* has a certain sort of conscious experiences, then any system with organization *F* has those experiences. But we cannot establish the logical necessity of the conclusion without establishing the logical necessity of the premise, and the premise is itself empirical. On the face of it, it is difficult to see why it should be logically necessary that *brains* with certain physical properties give rise to conscious experience. Perhaps the most tenable way to argue for this necessity is via a form of analytic functionalism; but in the context of using the Fading and Dancing Qualia arguments to *establish* this sort of functionalism, this strategy would be circular. It follows that the Fading and Dancing Qualia arguments are of little use in arguing for the logical and metaphysical impossibility of

absent and inverted qualia.

The arguments therefore fail to establish a strong form of functionalism upon which functional organization is *constitutive* of conscious experience; but they succeed in establishing a weaker form, on which functional organization *suffices* for conscious experience with natural necessity. We can call this view *nonreductive functionalism*, as it holds that conscious experience is determined by functional organization without necessarily being reducible to functional organization. As things stand, the view is just as compatible with certain forms of property dualism about experience as with certain forms of physicalism. Whether the view should be strengthened into a reductive version of functionalism is a matter that the Fading and Dancing Qualia arguments leave open.

In any case, the conclusion is a strong one. It tells us that systems that duplicate our functional organization will be conscious even if they are made of silicon, constructed out of water-pipes, or instantiated in an entire population. The arguments in this paper can thus be seen as offering support to some of the ambitions of artificial intelligence. The arguments also make progress in constraining the principles in virtue of which consciousness depends on the physical. If successful, they show that biochemical and other non-organizational properties are at best indirectly relevant to the instantiation of experience, relevant only insofar as they play a role in determining functional organization.

The principle of organizational invariance is not the last word in constructing a theory of conscious experience. There are many unanswered questions: we would like to know just what sort of organization gives rise to experience, and what sort of experience a given organization gives rise to. Further, the principle is not cast at the right level to be a truly *fundamental* theory of consciousness; eventually, we would like to construct a fundamental theory that has the principle as a consequence. In the meantime, the principle acts as a strong constraint on an ultimate theory.

References

Block, N. (1981). Troubles with functionalism. In (Block, ed.) *Readings in the Philosophy of Psychology, Volume 1*. Cambridge, MA: Harvard University Press.

Cuda, T. (1985). Against neural chauvinism. *Philosophical Studies*, 48, 111-27.

Horgan, T. (1984). Functionalism, qualia, and the inverted spectrum. *Philosophy and Phenomenological Research*, 44, 453-69.

Pylyshyn, Z. (1980). The 'causal power' of machines. *Behavioral and Brain Sciences*, 3, 442-4.

Savitt, S. (1982). Searle's demon and the brain simulator reply. *Behavioral and Brain Sciences*, 5, 342-3.

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3, 417-57.

What is the Unity of Consciousness?

Tim Bayne and David J. Chalmers

**Department of Philosophy
University of Arizona
Tucson, AZ 85721.**

1 Introduction

At any given time, a subject has a multiplicity of conscious experiences. A subject might simultaneously have visual experiences of a red book and a green tree, auditory experiences of birds singing, bodily sensations of a faint hunger and a sharp pain in the shoulder, the emotional experience of a certain melancholy, while having a stream of conscious thoughts about the nature of reality. These experiences are distinct from each other: a subject could experience the red book without the singing birds, and could experience the singing birds without the red book. But at the same time, the experiences seem to be tied together in a deep way. They seem to be *unified*, by being aspects of of a single encompassing state of consciousness.

This is a rough characterization of the unity of consciousness. There is some intuitive appeal to the idea that consciousness is unified, and to the idea that it must be unified. But as soon as the issue is raised, a number of questions immediately arise.

(1) *What is the unity of consciousness?* What does it mean to say that different states of consciousness are unified with each other, or that they are part of a single encompassing state? The idea of unity is multifaceted, and has been understood in many different ways by different thinkers. In some senses of "unity", the claim that consciousness is unified may be obvious or trivial; in other senses, the claim may be obviously false. So the first project in this area is to distinguish between varieties of unity, and to isolate those varieties that pose the most important questions.

(2) *Is consciousness necessarily unified?* Some thinkers (Descartes and Kant, for example) have argued that some sort of unity is a deep and essential feature of consciousness. On this view, the conscious states of a subject are necessarily unified: it is impossible for there to be a subject whose conscious states are disunified. On the other side, some thinkers (e.g. Nagel 1971) have argued that the unity of consciousness can break down. On this view, there are cases (especially neuropsychological cases, such as those involving patients with split brains) in which a subject's states of consciousness are disunified. Some (e.g. Dennett 1992) hold more strongly that consciousness is often or usually disunified, and that much of the apparent unity of consciousness is an illusion.

(3) *How can the unity of consciousness be explained?* If consciousness really is unified, and especially if it is necessarily unified, then it is natural to look for an explanation of this fact. What is it about

consciousness that yields this unity? Is unity a primitive feature of consciousness, or is it explained by something deeper? Further, the unity of consciousness may put strong constraints on a theory of consciousness. If consciousness is necessarily unified, then a correct theory of consciousness should at least be compatible with this unity, and we can hope that it will explain this unity.

We can see these three questions as clustering around the status of what we can call the unity thesis (UT):

Unity Thesis: Necessarily, any set of conscious states of a subject at a time is unified.

The first question raises the issue of how the notion of unity in the unity thesis is to be understood: what is it for a set of conscious states to be unified? The second question raises the issue of whether the unity thesis is true. The third question raises the issue of how, if the unity thesis is true, its truth might be explained.

In this paper we will address all three of these questions. Our central project will be to isolate a notion of unity on which the unity thesis is both substantive and plausible. That is, we aim to find a more precise version of the unity thesis that is neither trivially true nor obviously false. With such a thesis in hand, we will look at certain arguments that have been made *against* the unity of consciousness, to determine whether they are good arguments against the unity thesis as we understand it. And finally, after fleshing out the unity thesis further, we will apply the thesis to certain currently popular philosophical theories of consciousness, arguing that the thesis is incompatible with these theories: if the Unity Thesis is true, then these theories are false.

We will not aim to conclusively prove the unity thesis in this paper, and indeed we are not certain that it is true. But we aim to suggest at least that the thesis is plausible, that it captures a strong intuition about the nature of consciousness, and that there are no knockdown arguments against it. If the thesis is true, it is likely to have strong consequences for a theory of consciousness.

2 Varieties of Unity

To start with, we need to distinguish different notions of unity. In particular, we will distinguish various different ways in which different states of consciousness might be said to be unified with each other.

(1) Objectual unity. We can say that two states of consciousness are *objectually unified* when they are directed at the same object. For example, when I look at a red book, I have an experience of redness, and an experience of rectangularity. The color experience and the shape experience here are unified in a particularly strong way. They are present in my consciousness as directed at a single entity: the book. The same goes for my experience of a blue car moving down the street. Here I experience color, shape, and motion, all of which are unified by being directed at the same object. I might even have an auditory experience of the car's engine, and also experience this as directed at the same object. So there can be objectual unity across different sensory domains.

For two experiences to be objectually unified, their object need not actually exist. If I hallucinate a red book, then my experiences of redness and of rectangularity will be objectually unified, despite the book's nonexistence. On the other side of the coin, two experiences can be experiences of the same object without being objectually unified. I might see a car's shape and hear its noise, without anything in my conscious state tying the noise to the car (perhaps I perceive the noise as behind me, due to an odd environmental effect). If so, the experiences are not objectually unified. For objectual unity, what matters is that two states are experienced *as* being directed at a common object.

The notion of objectual unity is closely tied to a central issue in cognitive psychology and neurophysiology. When I look at a red square, the color and the shape may be represented in different parts of my visual system. But somehow these separate pieces of information are brought together so that I experience a single red square, so that I can identify and report a red square, and so on. This phenomenon is often referred to as *binding*, and the question of how it is achieved is often referred to as the *binding problem*. The binding problem is in large part the problem of how objectual unity is possible. As we will see, this divides into two problems in turn.

Objectual unity is an important phenomenon, but it will not be central for our purposes. Where objectual unity is concerned, the unity thesis is almost certainly false. While some sets of experiences are objectually unified with each other, it seems that most sets are not. For example, my experience of the color of the book and of the shape of the car are not objectually unified: they are experienced as being directed at *different* objects. My experiences of a bird singing and of a sharp pain do not seem to be directed at the same object at all. If so, then objectual unity cannot unify all of a subject's conscious states. For such a notion of unity, we must look elsewhere.

(2) Spatial unity. A related notion of unity is that of *spatial unity*. We can say that two conscious states are *spatially unified* when they represent objects as being part of the same space. For example, my experiences of a book and of a car are not directed at a common object, but they represent both objects as part of the same visual space. More generally, all of my visual experiences seem to be spatially unified in this way: every visual experience represents something spatially, and everything that is represented is represented as part of a common space. Auditory experiences usually represent objects as part of the same space; such auditory experiences are spatially unified with visual experiences.

The notion of being "represented as part of the same space" can be fleshed out in various ways, but the crucial idea will be something like this: a set of experiences are spatially unified if (i) each has spatial representational content, and (ii) the representational content of each is *comparable*, in the sense that the objects represented are represented as being in spatial relations to each other. So visual experience might represent a car as being near a tree, or behind a truck, or to the left of a building. Auditory experience might represent exhaust noise as coming from the same area as the car, or it might represent a siren as being much further away. This sort of comparability is endemic to visual experience and to much perceptual experience, and makes for a deep spatial unity in perception.

Like objectual unity, however, spatial unity does not yield a plausible version of the unity thesis. Some experiences seem to have no spatial representational content at all. An emotional experience such as that

of melancholy does not obviously represent anything as located within space. A conscious thought about philosophy might have no spatial content at all. If so, these conscious states are not spatially unified with other conscious states. As before, to find a notion of unity that unifies all of a subject's conscious states, we must look elsewhere.

(3) Subject unity. Let us say that two conscious states are *subject unified* when they are had by the same subject at the same time. So all of my current experiences — my perceptual experiences, my bodily sensations, my emotional experiences and conscious thoughts — are subject unified, simply because they are all *my* experiences.

If we construe the unity thesis as involving subject unity, it is certainly plausible. If a set of conscious states are had by a subject at a time, then they will be subject unified by definition. The trouble with this version of the unity thesis is that it is *trivial*. It is true by definition, and tells us nothing substantive about consciousness. As such, it cannot capture the intuition that there is some nontrivial way in which consciousness is unified. So subject unity will not be our central focus here.

Still, the notion of subject unity is at least useful in articulating the unity thesis. As it was characterized above, the unity thesis holds that if a set of experiences of a subject at a time is subject unified, then that set is unified. So in effect, the unity thesis states that subject unity entails unity. Now we simply need to find a notion of unity for which this entailment is both plausible and nontrivial.

(4) Subsumptive unity. We started the paper by invoking the intuition that there is some substantial sense in which *all* of a subject's experiences — including at least perceptual, bodily, emotional, and cognitive experiences — can all be unified. This sense is not object or spatial unity, since these notions do not apply to all of the relevant experiences. And this sense is not subject unity, since the resulting unity holds trivially. Rather, it is the sense that all of these experiences are somehow subsumed within a single state of consciousness.

We can say that two conscious states are *subsumptively unified* when they are both subsumed by a single state of consciousness. The notion of one state being subsumed by another should be taken as intuitive for now; we will spell it out more shortly. To take an example: it seems plausible that all of my visual experiences are subsumed by a single encompassing state of consciousness, corresponding to my visual field. More generally, my visual and auditory experiences might all be subsumed by a single encompassing state of perceptual consciousness. And it does not seem unreasonable to suppose that there is a single encompassing state of consciousness that subsumes all of my experiences: perceptual, bodily, emotional, cognitive, and any others.

We can think of this last encompassing state of consciousness, for a given subject, as the subject's *total* conscious state. When it exists, a subject's total conscious state might be thought of as the subject's conscious *field*. It can be thought of as involving at least a conjunction of each of many more specific conscious states: states of perceptual experience, bodily experience, emotional experience, and so on. But what is important, on the unity thesis, is that this total state is not *just* a conjunction of conscious states. It

is also a conscious state in its own right. If such a total conscious state exists, it can serve as the "singularity behind the multiplicity" — the single state of consciousness in which all of a subject's states of consciousness are subsumed.

It is worth pointing out certain sorts of unity with which subsumptive unity should not be confused. We are not talking about *gestalt* unity, where the conscious experiences of two different objects are deeply related in a way that transforms each of the experiences and produces a "gestalt" experience with a novel content. And we are not talking about *normative* unity, which requires some special coherence or consistency among multiple contents of consciousness. As we have characterized subsumptive unity, two conscious states might be subsumptively unified whether or not their contents stand in a special gestalt relation to each other, and whether or not they are especially consistent or coherent with one another. We are also not dealing with *neurophysiological* unity, which requires that conscious states involve a single area or mechanism in the brain. Finally, we are not dealing with *diachronic* unity, or the unity of consciousness across time. It might turn out that one or more of these notions is deeply related to the issues at hand, but none of them is our primary object of discussion.

To spell out the notion of subsumptive unity in more detail, we need to go into more detail about just what consciousness involves, and just what is involved in the idea of one conscious state being subsumed by another. This requires making some further distinctions.

3 Access unity and Phenomenal unity

What is it for a mental state to be a conscious state? There is no single answer to this question. As many have pointed out, the notion of "consciousness" is ambiguous, and is understood in different ways by different people. So to make progress we have to make distinctions. For our purposes, the most useful distinction is Ned Block's distinction between access consciousness and phenomenal consciousness (Block 1995).

A mental state is *access conscious* when a subject has a certain sort of access to the content of the state. More precisely, a state is access-conscious if by virtue of having the state, the content of the state is available for verbal report, for rational inference, and for the deliberate control of behavior. When I look at a red book, I can report the presence of the book ("there's a red book"), I can reason about it (e.g., concluding that I must have put it there when reading yesterday), and I can use its presence in deliberately directing my behavior (e.g., picking up the book and putting it back on the shelf). So my perception of the red book gives me the relevant sort of access to information about the red book. So my perceptual state here is access-conscious. One can also say that in such a case, the subject is access-conscious of the relevant object. So here, I am access-conscious of the red book.

In a similar way, many of my perceptual states are access-conscious, and so are many of my emotional and cognitive states. Not all mental states are access-conscious, however. In some cases, such as those involving subliminal perception, blindsight, or unconscious belief, a mental state represents information without that information being reportable or usable in rational control of reasoning and behavior. The

exact definition of access consciousness is somewhat flexible, and can be varied for different purposes. The most important point is that a state's being access-conscious is defined in terms of the causal role that the state plays within the cognitive system, and in particular in terms of the role that the state plays in making information available to other parts of the system.

A mental state is *phenomenally conscious* when there is something it is like to be in that state. When a state is phenomenally conscious, being in that state involves some sort of subjective experience. There is something it is like for me to see the red book — I have a visual experience of the book — so my perception of the book is phenomenally conscious. There is something it is like to hear the bird singing, and to feel the pain in my shoulder, so these states are phenomenally conscious. There is something it is like to feel melancholy, and there is arguably something it is like when I think about philosophy. If so, then these states are phenomenally conscious. Phenomenal consciousness is often taken to be the most important sort of consciousness, and to be the sort of consciousness that poses the most difficulty for scientific explanation.

There is a close empirical connection between phenomenal consciousness and access consciousness. It is arguable that the two almost always go together empirically: when a state is phenomenally conscious, it is access-conscious, and vice versa. That is, when there is something it is like to be in a state, a subject can usually report the contents of the state and use it to directly guide reasoning and behavior. And when a subject can report the contents of a state and use it to directly guide reasoning and behavior, there is usually something it is like to be in that state. So when I am phenomenally conscious of the red book, I am access-conscious of it, and vice versa.

Despite this empirical connection, there is plausibly a *conceptual* distinction between access consciousness and phenomenal consciousness. Access consciousness is defined in terms of the causal role that a state plays, whereas phenomenal consciousness is defined in terms of the way the state feels. It is arguable that we can at least *imagine* states that are access-conscious without corresponding states of phenomenal consciousness (the philosophers' zombie, which is functionally like a normal human being but without any conscious experience, would be one such imaginary case). And it seems that we can know about another being's states of access consciousness without knowing about their states of phenomenal consciousness: one might know what information is available for report and for behavioral control in a cognitive system without being in a position to know what it is like to be that system.

When there is something it is like to have a mental state, we can say that the mental state has a phenomenology, or a phenomenal character. Slightly more formally, we can say that such mental states have *phenomenal properties*, or *qualia*, which characterize what it is like to be in them. We can also say that *subjects* have phenomenal properties, characterizing aspects of what it is like to be a subject at a given time. We can then say that a phenomenal state is an instantiation of such a property. For example, the state of experiencing a certain sort of reddish quality is a phenomenal state.

When a subject is in a phenomenally conscious mental state, the subject will thereby be in a phenomenal state that reflects the phenomenology of being in the mental state. For example, if there is something it is like for a given subject to believe that Paris is in France, the subject will be in a corresponding

phenomenal state. But the phenomenally conscious mental state and the phenomenal state may be distinct states. For example, it may be that there is now a certain phenomenology associated with my occurrent belief that P, while it is also possible for me to believe that P with a different phenomenology or no phenomenology at all. If so, the belief state (this instance of believing such-and-such) is a phenomenally conscious mental state, but it is not a phenomenal state. There is a special class of phenomenally conscious mental states such that the mental state and the corresponding phenomenal states are identical: phenomenal states themselves. Phenomenal states are at the core of phenomenal consciousness.

We can use the distinction between access consciousness and phenomenal consciousness to make a distinction between two corresponding notions of unity: access unity and phenomenal unity. Broadly speaking, two conscious states are *access-unified* when they are jointly accessible: that is, when the subject has access to the contents of both states at once. Two conscious states are *phenomenally unified* when they are jointly experienced: when there is something unified it is like to be in both states at once.

We can construct more precise versions of phenomenal and access unity by combining these distinctions with the distinctions outlined earlier between objectual unity, spatial unity, and field unity. These distinctions crossclassify each other, so that one can isolate notions of objectual phenomenal unity, objectual access unity, spatial access unity, subsymptomatic phenomenal unity, and so on. The distinction applies less clearly to the notion of subject unity, so we will set that notion aside here.

We can say that two conscious states are *objectually access-unified* when their contents involving attributing properties to a single object of representation, and when these contents are jointly accessible within the system. The contents will be jointly accessible when their *conjunction* is available for report and the rational control of reasoning and behavior. When I am conscious of a red square, I can report the presence of red, and the presence of a square, but I can also report the presence of a red square. Similarly, the presence of a red square can be used in guiding my reasoning and my behavior. So my perception of red and my perception of a square are not just individually access conscious: they are access-unified.

We can say that two conscious states are *objectually phenomenally unified* when they are experienced as representing a single object. When I am conscious of a red square, I experience the presence of red and I experience the presence of a square, but I also experience the presence of a red square. There is a distinctive sort of unity involved in what it is like to experience the redness and the squareness simultaneously here: the two states are unified by being experienced as aspects of a single object.

Objectual access unity and objectual phenomenal unity correspond to two distinct aspects of the binding problem. It has often been pointed out that there are actually two binding problems (see e.g. Revonsuo 1999). The first is that of how a system such as the brain manages to bring together two separately represented pieces of information (e.g., representations of color and shape in different areas of the visual cortex), so that these can play a joint role in the control of behavior (e.g. so that we can report the presence of a red square and a blue circle, rather than a red circle and a blue square). This is a sort of engineering problem concerning the design of the cognitive system; one can think of it as the neurophysiological or cognitive binding problem. This binding problem is the problem of explaining

objectual access unity. The second binding problem is that of explaining how it is that we perceptually experience separate pieces of information as bound together in pertaining to the same object. This is the problem of explaining objectual phenomenal unity. On the face of it, these two problems are distinct: one could solve the neurophysiological binding problem, giving an explanation of how two pieces of information are brought together in the brain to be jointly accessible, while still having no explanation of why the jointly accessible information should be experienced. So objectual phenomenal unity and objectual access unity are at least conceptually distinct.

One can make a similar distinction between spatial phenomenal unity and spatial access unity. We can say that two conscious states are spatially access-unified when they have spatial representational contents that can be jointly accessed by the cognitive system, so that they can be spatially compared and so that the results of the comparison can be made available for report, reasoning, and behavioral control. For example, when I see a car and a tree, I do not just have access to their spatial locations individually; I also have access to the spatial locations jointly, in that I can report that the car is to the left of the tree. So these two perceptual states are spatially access-unified. Two conscious states are spatially phenomenally unified when they involve experiencing entities as part of the same space; as part of the same phenomenal space, one might say. I experience the car as being in the same space as the tree, and to the left of it; so these two states are spatially phenomenally unified.

The most important distinction is that between subsumptive phenomenal unity and subsumptive access unity. These notions apply to two arbitrary conscious states, as long as they are phenomenally conscious in the first instance and access-conscious in the second. Because these are the most important versions of unity, we will henceforth usually speak simply of "phenomenal unity" and "access unity", where it is understood that we are referring to phenomenal field unity and access field unity respectively.

We can say that two conscious states are *subsumptively access-unified* (or simply *access-unified*) if the *conjunction* of their contents is available for verbal report, reasoning, and the deliberate control of behavior. So if mental state A has content P and mental state B has content Q, these states will be individually access-conscious if the information that P is available for report and for control, and if the information that Q is available for report and control. They will be *jointly* access-conscious, or they will be access-unified, if the information that $P \wedge Q$ is available for report and control. More briefly: two states A and B are access-unified if and only if the subject is access-conscious of the conjunction of their contents. In this case, there is an access-conscious mental state with the conjunctive content: this conjunctive mental state can be seen as subsuming the original states A and B.

For example, when I see a book and feel a pain, I can report the presence of the book and of the pain individually, but I can do more than that: I can report them simultaneously. I can also reason about the book and the pain jointly, and use information about both to jointly control my behavior (e.g. looking in the book for a remedy for the pain, or ceasing to read the book to help alleviate the pain). Because of the accessibility of this conjunctive content, the two states are (subsumptively) access-unified. Similarly, I can often jointly report or reason about an emotion and a sound: if so, the emotional state and the auditory state are access-unified. And so on.

It is worth noting that for a state to be access-conscious, it is not required that the content of the state actually be *accessed*, in the sense that it is directly used for report or for control. What matters is that it be *accessible*, in a certain direct sense, or "poised" for use in access, as Block puts it. The same goes for access unity. For two states to be access-unified, they need not be simultaneously accessed at any given moment. What matters is that they are simultaneously accessible, in that it would be possible for a subject to jointly report them, and to use them jointly in reasoning and behavior control. Typically, our conscious states are not jointly accessed, but they are much more often jointly accessible. It is joint accessibility that matters for our notion of unity.

We can use the notion of access unity to put forward a version of the unity thesis:

Access Unity Thesis: Necessarily, any set of access-conscious states of a subject at a time is access-unified.

This thesis appeals to the notion of a set of states being access-unified. This is a natural generalization of the notion of two states being access-unified. We can say that a set of states is access-unified if the contents of all of the states are jointly accessible.

It might be objected that in requiring that *any* set of a subject's access-conscious states is access-unified, the thesis is highly implausible. A subject might have a large (possibly infinite) number of access-conscious states, and the conjunction of the contents of these states might be so complex that it is implausible that a subject could have access to this conjunction. The full conjunction would not be reportable or directly available to guide reasoning and behavior. To get around this, we could put forward a slightly weakened version of the thesis:

Pairwise Access Unity Thesis: Necessary, any two access-conscious states of a subject at a time are access-unified.

One might argue that the pairwise version is too weak to count as a full unity thesis (which requires unity of all states at a time), or that it suffers from the same problems as the full unity thesis (since it entails that conjunctions of conjunctions will be access-conscious, and so on). But none of this will matter for our purposes, since as we will argue shortly, even a weak version of the access unity thesis, limited to pairwise unity of relatively simple access-conscious states, is straightforwardly false.

We can say that two conscious states are *subsumptively phenomenally unified* (or simply *phenomenally unified*) if there is something it is like for a subject to be in both states simultaneously. That is, two states are phenomenally unified when they have a *conjoint phenomenology*: a phenomenology of having both states at once that subsumes the phenomenology of the individual states. When A and B are phenomenally conscious states, there is something it is like for a subject to have A, and there is something it is like for a subject to have B. When A and B are phenomenally unified, there is not just something it is like to have each state individually: there is something it is like to have A and B together. And the phenomenology of being in A and B together will carry with it the phenomenology of being in A

and the phenomenology of being in B.

For example, when I look at the book while feeling a pain, there is something it is like to see the book (yielding a phenomenal state A), and there is something it is like to feel the pain (yielding a phenomenal state B). But there is more than this: there is something it is like to see the book while feeling the pain. Here there is a sort of conjoint phenomenology, that carries with it the phenomenology of seeing the book, and the phenomenology of feeling the pain. As in the discussion of field unity, we can think of the conjoint state here as involving at least the conjunction A&B of the original phenomenal states A and B. But importantly, the conjoint state is itself a phenomenal state: a single complex state of consciousness that subsumes the individual states of consciousness A and B. It is this encompassing state of consciousness that unifies A and B.

More generally, we can say that a set of conscious states is phenomenally unified if there is something it is like for a subject to have all the members of the set at once, and if this phenomenology subsumes the phenomenology of the individual states. As a special case, we can say that the set consisting of all of a subject's conscious states at a given time is phenomenally unified if there is something it is like for the subject to have all these states at once, where this phenomenology subsumes the phenomenology of the individual states. If so, then the subject has a *total phenomenal state* that encompasses all of the subject's phenomenal states. One can think of a total phenomenal state as capturing what it is like to be a subject at a time. If a subject has a total phenomenal state, there is a clear sense in which all of a subject's phenomenal states are unified within it.

We can put forward a phenomenal version of the unity thesis, as follows:

Phenomenal Unity Thesis: Necessarily, any set of phenomenal states of a subject at a time is phenomenally unified.

This is not quite the same as the unity thesis that falls out of our original definition. Such a thesis would hold that any set of *phenomenally conscious mental states* of a subject at a time will be phenomenally unified. But the two theses are clearly equivalent. The first version (regarding phenomenally conscious mental states) entails the second version (regarding phenomenal states) as a special case. In reverse, the second version entails that for any set of phenomenally conscious mental states, their associated phenomenal states will be phenomenally unified. So there will be a phenomenal state that subsumes each of the original phenomenal states. So there will be something it is like to be in all the original mental states simultaneously that subsumes what it is like to be in them individually. So the original mental states will be phenomenally unified, and the first version of the unity thesis holds.

One can also put forward slightly weaker versions of the phenomenal unity thesis:

Pairwise Phenomenal Unity Thesis: Necessarily, any two phenomenal states of a subject at a time are phenomenally unified.

Total Phenomenal Unity Thesis: Necessarily, the set of all phenomenal states of a subject at a time is phenomenally unified.

The original phenomenal unity thesis clearly entails the pairwise unity thesis and the total unity thesis. The pairwise thesis does not obviously entail the first version. It is plausible that subsumption is transitive, so that necessarily, if A subsumes B and B subsumes C, then A subsumes C. If so, the pairwise unity thesis will entail the phenomenal unity thesis for any finite set of phenomenal states, as any pair of these will be subsumed by a single phenomenal state, and any pair of those in terms will be subsumed by a single phenomenal state, and so on. But the pairwise thesis does not obviously entail the original thesis where infinite sets of phenomenal states are concerned. The total unity thesis entails the original phenomenal unity thesis, however: if there is a state that subsumes each phenomenal state of the subject, that state will also subsume each member of an arbitrary set of phenomenal states of the subject, so that set will be phenomenally unified. So the total unity thesis and the original phenomenal unity thesis are equivalent.

The total unity thesis arguably captures the central intuition behind the unity of consciousness. This thesis suggests that there is always a single phenomenal state that subsumes all of the phenomenal states of a subject at a time. That is, it suggests that any conscious subject at any time has a total phenomenal state. If a subject has a total phenomenal state, subsuming every specific phenomenal state of the subject, then the subject's consciousness will be unified in a deep way.

It might be objected that when a subject experiences a number of phenomenal states at once, the original phenomenal states will be transformed. For example, it might be phenomenally different to see a red book in the context of a moving car than to see a red book on its own, and the phenomenal state that was present when one saw the book on its own might not be present at all. This may be so, but it is no objection to the unity thesis. The unity thesis says that the phenomenal states had by a subject *at a time* are subsumed by a complex phenomenal state. So the experience of a red book and a moving car at a given time should subsume the experience of the red book at that time and the experience of the moving car at that time. It is not required that the complex experience should subsume the experience of a red book as the subject might have it at a different time, in a different context. If the experience of the book is itself transformed by the context of the car, then it is the transformed experience that will be subsumed by the complex state.

It might also be objected that these unity theses are *trivial*. If a subject has a set of phenomenal states, there will automatically be a phenomenal state that subsumes them: the conjunction of the original states. But this is not a trivial claim. It is trivial that if a subject is in a number of phenomenal states, the subject will be in the conjunction of those states. But it is nontrivial that this conjunction will itself be, or be subsumed by, a phenomenal state. That is, it is nontrivial that there will be *something it is like* to be in the conjunctive state. This can be seen from the fact that some philosophers deny the total unity thesis, or at least entertain its denial. For example, when Hurley (1998) discusses the possibility that the unity of consciousness could break down and that consciousness could be "partially unified" (so that two phenomenal states are each unified with a third state, but not with each other), she says:

Therefore, we cannot imagine what it is like for there to be partial unity. That doesn't show partial unity is unintelligible, because being partially unified isn't the sort of thing there *could be* anything it is like to be. We shouldn't expect to be able to imagine what it is like.

In general, it seems that a case in which the unity of consciousness breaks down would be precisely a case in which there is no total phenomenal state of the subject: that is, there is nothing it is like to be the subject at that time, or at least there is no single something-it-is-like that captures all the phenomenal states of the subject. Such a subject would have states with a local phenomenal character, but there would be no global phenomenal character involved in having these states. It is certainly very hard to see how this could be the case. Indeed, one might suspect (as we do) that such a scenario is impossible and perhaps incoherent. But to say this is not to say that the unity thesis is trivial: it is a substantive thesis about the nature of consciousness. This is reflected by the fact that (as we discuss later in the paper) certain theories of consciousness entail that the unity thesis is false. If so, then the thesis puts substantive constraints on a theory of consciousness.

4 When Access Unity Breaks Down

The access unity thesis holds that necessarily, any two access-conscious states are access-unified. This entails that whenever a subject is access-conscious of P and is access-conscious of Q , the subject will be access-conscious of $P \wedge Q$. This thesis is clearly false.

To see that the thesis is false, we need only note that it is possible for a subject to be access-conscious of P and access-conscious of Q , without being access-conscious of $P \wedge Q$. For this to happen, it should be the case that P is reportable and available for guiding reasoning and behavior, and that Q is reportable and available for guiding reasoning and behavior, but that $P \wedge Q$ is not reportable and not available for guiding reasoning and behavior. This can happen in a quite straightforward way. All that is required is that there be an *access bottleneck*. This will be a pathway of information access through which only a limited amount of information can pass at one time. If P and Q are both accessible only through the bottleneck, and if each carry an amount of information that is near the capacity of the bottleneck, then P and Q will be individually accessible, but the conjunctive content $P \wedge Q$ will not.

This is not merely a hypothetical description of an imaginary case. Such access bottlenecks can occur in real cognitive systems, and are revealed by a number of experiments in the psychological literature. Perhaps the clearest example of such a bottleneck is given by a famous experiment by George Sperling (1960).

In Sperling's experiment, a subject is presented with a matrix consisting of three rows with four letters each. The matrix is flashed only briefly, for 250 milliseconds. After the matrix vanishes, a tone sounds, indicating whether the subject is to report the contents of the first, second, or third row. When subjects are required to report the contents of the top row, on average they correctly report 3.3 of the four letters in that row. The same goes when they are required to report the contents of the middle row, or of the bottom row. But when subjects are asked to report the contents of the entire matrix, on average they

correctly report 4.5 of the twelve letters. So, to simplify a little, it seems that the subject has access to the information in any single row, but the subject does not have joint access to the information in all three rows.

In this case, it is natural to hold that the subject (just after the matrix disappears, before the tone sounds) is access-conscious of the contents of any individual row. Recalling that access consciousness requires accessibility for report and for reasoning and behavior: the contents of each row are available for report (individually), and could presumably be used to guide reasoning about those contents and to guide behavior. But it also seems that the subject is not access-conscious of the conjunctive contents of the whole matrix, or of any two rows. The conjunctive contents of more than one row are not available for verbal report, and presumably are not available to guide reasoning and behavior. If so, then a subject can be access-conscious of P (one row) and of Q (another row), without being access-conscious of $P \wedge Q$ (both rows). So two access-conscious states of a subject at a time can fail to be access-unified, and the access unity thesis is false.

We do not claim that the Sperling experiment alone *proves* that the access unity thesis is false. There are other possible interpretations of the experiment: for example, one could hold that the subject has some sort of internal access to the conjunctive content, but that the process of report destroys this access. But the interpretation we have suggested is a natural one: on the face of it, the conjunctive content does not seem to be available for any sort of reasoning or control, although the individual contents are available taken one at a time. And importantly, whether or not this interpretation is correct of the actual case, it seems to be a perfectly coherent interpretation, one that describes a perfectly reasonable way for a cognitive system to function.

Indeed, given a natural design for cognitive systems with limited resources, we would expect certain restrictions on the flow of information in access and control, and we would expect access bottlenecks to arise in some cases. It may be that most of the time, when a subject has access to P and to Q, the subject has access to $P \wedge Q$. But this sort of joint access clearly cannot hold *necessarily*. So even if there is a reasonably high degree of access unity in ordinary conscious states, this sort of access unity cannot hold across the board.

This breakdown of access unity does not entail a breakdown of phenomenal unity. This can be seen by examining the Sperling case. It is difficult to know exactly what is going on in the phenomenology of the subject who is undergoing the Sperling experiment, before being asked about the contents of a row. Perhaps the details of all nine letters are present in the subject's phenomenology (as some subjects report); perhaps these details are not present, and there is merely an indeterminate patch in each cell of the matrix; or perhaps there is something in between. But whatever the exact phenomenology here, there is little reason to suppose that phenomenal unity breaks down.

No matter what it is like for a subject to experience each individual cell of the matrix in the Sperling case, it is plausible that there will be something it is like for the subject to see the entire matrix. And it is plausible that the phenomenology of seeing the matrix will subsume the phenomenology of seeing the

individual cells. If the phenomenology of seeing a cell involves just a hazy patch, then the phenomenology of seeing the matrix will plausibly involve nine hazy patches. If the phenomenology of seeing a cell involves a detailed shape, then the phenomenology of seeing the matrix will plausibly involve nine detailed shapes. Either way, the individual phenomenal states are subsumed by the overall phenomenal state. So there is no reason to deny phenomenal unity here.

At most, this sort of case suggests that a subject does not always have simultaneous access to the contents of all of the subject's phenomenal states. If the subject is indeed experiencing the details of all nine letters, then the subject is in a position where the contents of these experiences can be accessed and reported only a few at a time, and not all at the same time. There is nothing paradoxical or contradictory about this. It simply suggests that a subject's access to a total phenomenal state is sometimes piecemeal. But this is just what we might expect.

One consequence of this is that access consciousness and phenomenal consciousness can come apart. We have seen that the subject is access-conscious of the individual letters but not of their conjunction. And it is natural to hold that either (i) the subject is phenomenally conscious of neither the individual letters nor their conjunctions, or (ii) the subject is phenomenally conscious of both the individual letters and their conjunction. In case (i), a subject is access-conscious of an individual letter but not phenomenally conscious of it. In case (ii), a subject is phenomenally conscious of the conjunction but not access-conscious of it. Either way, access consciousness and phenomenal consciousness of a given content can come apart. Our own view is that description (ii) is somewhat more plausible. If this is so, we can still hold that access consciousness and phenomenal consciousness are correlated with each other for *simple* contents. But access consciousness and phenomenal consciousness will not always be correlated for *complex* contents.

The moral of all this is that a breakdown of access unity does not entail a breakdown in phenomenal unity. There is a sense in which a breakdown of access unity is a "disunity" in consciousness, but it is a relatively shallow sense. Such a breakdown is quite compatible with an underlying phenomenal unity. Of course we have not demonstrated that no breakdowns of access unity involves a breakdown of phenomenal unity. But this discussion does strongly suggest that one cannot *infer* a breakdown of phenomenal unity from a breakdown in access unity. To accept a breakdown of phenomenal unity, one would need some quite distinct reason.

An opponent might try to argue that the Sperling case is a case where phenomenal unity breaks down. For example, the opponent might argue that the phenomenology of seeing each individual cell involves a detailed letter, but that the phenomenology of seeing the whole matrix does not, and that any "global" phenomenology here involves only hazy patches. Such a response would seem unmotivated and implausible, on the face of it, at least in the absence of much supporting argument. If the phenomenology of each letter is detailed, then there seems to be good reason to hold that this phenomenology is present in a global phenomenal state. And even if it is *coherent* for an opponent to hold this, it is equally coherent (and seemingly more plausible) to deny this, and to hold that the experience of the letters is phenomenally unified. The mere coherence of the denial is enough to show that one cannot *infer* a breakdown in phenomenal unity from a breakdown of access unity.

5 Can Phenomenal Unity Break Down?

We think that there is a strong *prima facie* case that the unity thesis is true. This *prima facie* case is brought out by the fact that there seems to be something *inconceivable* about phenomenal disunity. It is difficult or impossible to imagine a subject having two phenomenal states simultaneously, without there being a conjoint phenomenology for both states. And there is a sense that there is something incoherent about the suggestion. This *prima facie* inconceivability — whether it takes the form of unimaginability or apparent incoherence — gives at least some reason to believe that cases in which phenomenal unity break down are impossible, so that the unity thesis is true.

But this is only a *prima facie* case. There are some possible scenarios that humans cannot imagine, and there are arguably some possible scenarios that no being could imagine. And the judgment of incoherence in this case is not so strong that it could not be incorrect. So the *prima facie* case for the unity thesis needs to be balanced with the case *against* the unity thesis. A number of philosophers and scientists have argued that the unity of consciousness can break down. So to assess the unity thesis, one needs to examine these arguments in order to see what force they have against the unity thesis as we have understood it.

By far the most common reason for holding that the unity of consciousness can break down is grounded in neuropsychology. It is widely held that patients in various unusual neuropsychological states have a disunified consciousness. The paradigm case here is that of a *split-brain* patient, whose corpus callosum has been severed for medical purposes, preventing the left and right hemispheres of the cerebral cortex from communicating directly (although there is still some connection through lower areas of the brain). Such a patient behaves in a surprisingly normal fashion much of the time, but in certain circumstances they behave quite unusually. For example, when presented with different pictures in different halves of their visual field (e.g., a cat on the left and a dog on the right), and asked to report the contents, the patient will report seeing only a dog, since the left hemisphere, which dominates speech, receives input from the right visual field. When asked to write down what they see with their left hand (which is controlled by the right hemisphere), such a patient may slowly write "CAT"; with the right hand, the patient may write "DOG". If a patient writes with her left hand in her right visual field, a conflict may occur when the patient sees what is written, and in some cases the right hand scratches out what the left hand has written.

It is often held that in cases like this, consciousness is disunified. On one interpretation (e.g. Puccetti 1981), there are two distinct subjects of consciousness, corresponding to each hemisphere. Such an interpretation is actually compatible with the unity thesis, since the unity thesis requires only that every subject have a unified consciousness. More threatening to the unity thesis are interpretations on which there is a single subject with a disunified consciousness. Some (e.g. Marks 1980) hold that the subject has two separate streams of consciousness, at least under experimental conditions. Others (e.g. Lockwood 1989) hold that the subject has a fragmented consciousness with nontransitive unity between the states: for example, the experiences of "CAT" and of "DOG" might each be unified with some background

emotional state, but not with each other. Others (e.g. Nagel 1971) hold that our conceptual framework in speaking of subjects may simply break down in this area.

Adjudicating this question requires a very detailed examination of both the empirical details and the philosophical analysis of these phenomena, which we cannot provide here. Here, we will simply note that given what we have said so far in this paper, the advocate of a phenomenal unity thesis has a natural line of response.

It is plausible that in split-brain cases, there is some sort of breakdown of *access* unity. If we assume that there is a single subject, then it seems that the subject in the case above has at least a weak sort of access both to the presence of a cat and to the presence of a dog, and can use each in reasoning and in the control of behavior. But it seems that the subject has no access to a conjunctive content involving both the cat and the dog. The conjunctive content is not reported, and plays no apparent role in reasoning and in the control of behavior. So this may well be a case in which access unity fails. In this case, it seems that two accessed contents are not jointly accessible, because of a disconnection between the relevant access mechanisms.

But as we have seen, a breakdown of access unity does not entail a breakdown of phenomenal unity. So the possibility remains open that split-brain subjects have a unified phenomenal field, with some sort of conjoint phenomenology subsuming each of the separate contents. It is just that the subject has pathologies of access, so that the contents of the field are accessible only singly and not jointly. If so, the subject in the experiment described has a phenomenal field that includes experiences of both "CAT" and "DOG". The subject simply has no conjoint access to these contents. Of course this implies that the subject has highly imperfect knowledge of her conscious states: she will believe (in both "halves" of the brain) that she is experiencing only one word, when in fact she is experiencing two. But it is plausible for many other reasons that knowledge of consciousness is fallible, and it is not unreasonable to suppose that in cases of brain damage, this fallibility might be quite striking.

Of course nothing here proves that this interpretation is correct. It does suggest, however, that we should not be too quick to conclude that these cases involve a breakdown of phenomenal unity. Most of those who have discussed these cases have not carefully distinguished the relevant notions of unity and consciousness (an exception is Marcel 1994, who distinguishes "reflexive consciousness" from "phenomenal experience" and argues that the disunity concerns the former), and have often discussed things in terms of access and related functional notions. Once we distinguish access unity from phenomenal unity, it becomes clear that the direct evidence concerns access disunity, not phenomenal disunity. To establish phenomenal disunity requires substantial further argument. It may be that such arguments can be given, but the case is far from clear.

One might say something similar about other disorders of consciousness, such as dissociative identity disorder (multiple personality disorder). In this case, it seems that there are pathologies of access between different parts of a cognitive system. But it seems quite tenable to hold that nevertheless, there is a single field of consciousness at any given time, subsuming the conscious states of the subject, even if they are in certain respects mutually inaccessible. Of course as in the split-brain case, the subject may well have

various false beliefs about her own consciousness (e.g. that the various states belong to different subjects), but again this is not unexpected. Indeed, there are other cases involving "alien thoughts" in schizophrenic patients where it seems most plausible to hold that a subject incorrectly attributes some of her conscious states to a subject other than herself.

To completely assess this thesis requires much further analysis. But for now, we conclude that the empirical case against the phenomenal unity thesis is at best inconclusive. Given the strong *prima facie* positive case for accepting the phenomenal unity thesis, this suggests that the unity thesis remains quite plausible.

6 Formalizing the unity thesis

6.1 Subsumption and entailment

For further analysis, we need to clarify the phenomenal unity thesis, and the corresponding notion of phenomenal unity. We have said that a set of states is phenomenally unified when there is something it is like to be in all those states at once. When this is the case, the subject will have a phenomenal state (corresponding to the conjoint what-it-is-like) that *subsumes* each of the states in the original set. So phenomenal unity can be seen as a sort of subsumptive unity, and the phenomenal unity thesis on the table is a sort of subsumptive unity thesis.

Subsumptive Unity Thesis: For any set of phenomenal states of a subject at a time, the subject has a phenomenal state that subsumes each of the states in that set.

There are also closely related total and pairwise subsumptive unity theses, requiring subsumptive unity only for pairs of phenomenal states or only for the complete set of a subject's phenomenal states at a time, but we can focus on the thesis above for now.

As it stands, the notion of subsumption is something of an intuitive primitive. There are some things we can say about it. It is a relation among token phenomenal states. It is plausibly reflexive (a state subsumes itself), antisymmetric (if A subsumes B and B subsumes A, then $A = B$), and transitive (if A subsumes B and B subsumes C, then A subsumes C). Note that reflexivity eliminates any apparent problem of regress in the unity thesis (if A and B are subsumed by C, there is no need for a further state to subsume A and C, since C subsumes itself).

The paradigm case of subsumption is the relation between a complex phenomenal state and a simpler state that is intuitively one of its "components". One might think of subsumption as analogous to a sort of mereological part/whole relation among phenomenal states, although this should be taken as an aid to intuition rather than as a serious ontological proposal, at least at this point. It is also useful to stipulate that subsumption holds between a phenomenal state and less specific states that intuitively correspond to the same experience: for example, that the state of experiencing a sharp pain subsumes the corresponding

state of experiencing a pain. This sort of subsumption is required in order for it to be possible that a highly specific total phenomenal state can subsume all of a subject's phenomenal states, including unspecific states.

It should be noted that there are alternatives to analyzing phenomenal unity in terms of subsumption. Often, phenomenal unity is analyzed in terms of an intuitive relation of co-consciousness, where this relation is taken as primitive. We think that the analysis in terms of subsumption runs deeper in certain respects than a primitive analysis in terms of co-consciousness, and offers the promise of further analytic tools, as discussed below. But the exact relation between these notions is an open question. (Dainton (2000) gives a thorough and insightful analysis of the unity of consciousness in terms of a primitive co-consciousness relation; Bayne (2001) discusses the relationship between the different accounts.)

The notion of subsumption is connected to the notion of "what it is like" in at least the following sense: when A subsumes B, what it is like to have B is an aspect of what it is like to have A. Of course this appeals to the unexplained notion of an "aspect". One might try to go further by *defining* subsumption wholly in terms of notion of "what it is like" as follows: A phenomenal state A subsumes phenomenal state B when what it is like to have A and B simultaneously is the same as what it is like to have A. This seems to capture the connection articulated above, and it also can ground the connection between subsumptive unity and the original definition of phenomenal unity. If there is something it is like to be in a set of states (as the original definition requires), then this phenomenology will correspond to a phenomenal state A of the subject, and it is clear that this state will subsume the states in the original set in the sense defined above. It is arguable that the defined notion of subsumption goes beyond the intuitive notion in certain respects (someone might hold that the "what it is like" locution can be read such that what it is like to have A and B differs from what it is like to have A, even when A subsumes B), and we will not rely on it in what follows, but nevertheless it can serve as a useful aid to the understanding.

(Extending this line of thought, one could say that a state A *precisely subsumes* a set of states S what what it is like to be in A is the same as what it is like to simultaneously be in the members of S. Then if A precisely subsumes S, A subsumes each of the members of S, but the reverse entailment does not hold. For example, a subject's total state of consciousness subsumes each of the subject's visual experiences, but it does not precisely subsume the set of them. One could then articulate a *Correspondence Thesis* holding that for any set of phenomenal states of a subject at a time, there is a corresponding phenomenal state that precisely subsumes that set. The correspondence thesis is formally stronger than the original subsumptive unity thesis: the existence of a total phenomenal state suffices for the truth of the original thesis, but it does not suffice for the truth of the correspondence thesis. The correspondence thesis nevertheless has some intuitive plausibility, and one could argue that this thesis, rather than the subsumptive unity thesis, best captures the idea articulated in the original phenomenal unity thesis. The difference between these theses will not be important for our purposes, however.)

There is a close relation between subsumption and *entailment*. Let us say that a state P entails a state Q when it is impossible (logically or metaphysically impossible) for a subject to instantiate P without instantiating Q. Then it seems clear that when a phenomenal state P subsumes a phenomenal state Q, P will entail Q. For example, if P involves the phenomenal character as of seeing a red book and hearing a

bird singing, and if Q involves the phenomenal character as of seeing a red book, then it is impossible to have P without having Q. The same goes with any case of subsumption: by its nature, the subsuming state carries with it the subsumed state.

Note that strictly speaking, entailment is a relation among state-types, while subsumption is a relation among state-tokens. For present purposes, we can regard entailment as derivatively a relation among state-tokens, so that one state-token entails another when there is entailment between the corresponding state-types (although see below). We will generally pass over this nicety in discussion, acknowledging it where it is relevant. Note also that a phenomenal state A entails a phenomenal state B if necessarily, a subject in A is also in B — *not* if the content of A entails the content of B.

The close relation between entailment and subsumption raises an interesting possibility: perhaps we can simply *define* subsumption in terms of entailment? That is, perhaps we can hold that phenomenal state A subsumes B when A entails B? If this were possible, instead of relying on a novel primitive relation, we could analyze unity in terms of a well-understood relation that allows the use of standard logical tools. To help assess this possibility, we can define a corresponding notion of unity and a corresponding unity thesis:

A set of phenomenal states of a subject at a time is *logically unified* when the subject has a phenomenal state that entails each of the phenomenal states in that set.

Logical Unity Thesis: For any set of phenomenal states of a subject at a time, the subject has a phenomenal state that entails each of the states in the set.

This gives an attractively simple formulation of the unity thesis, and one that has some intuitive force. Unfortunately there is an obstacle to replacing subsumption by entailment. We know that when A subsumes B, A entails B. But the reverse is not obviously the case. In fact, there are two ways in which it may seem that A could entail B without subsuming B.

First, A and B might correspond to intuitively distinct experiences that share a type. For example, a subject might have two pains at the same time, or two experiences of red, and so will have two distinct phenomenal states of the same type. In this case, one state-type will entail the other, so if entailment among tokens is derivative on entailment among types, one state-token will entail the other. In this case, it is not plausible to hold that one state subsumes the other. (What it is like to have A and B simultaneously is quite different from what it is like to have A.) One might instead refine the definition of entailment among state-tokens, requiring that it is impossible for one token to exist without the other, in addition to the requirement that one type cannot exist without the other. But one can also deal with this case by a strategy discussed below.

Second, A and B could be intuitively distinct phenomenal states that do not share any simple type, but that are nevertheless necessarily connected. This would involve a sort of *gestalt unity*, which involves constraints on the co-occurrence of distinct phenomenal states. For example, perhaps there are cases

where feeling a pain in one's shoulder while also experiencing a splitting headache produces a unique sort of pain that could not be experienced in the absence of the headache. Or perhaps seeing a certain person in the middle of a crowd produces a unique sort of visual experience of that person that could not be had in the absence of the experience of the crowd. Or perhaps (to use an example from Dainton 2000) the experience of the boundaries of a Kanisza triangle is of a special sort that could not be had in the absence of the circles in which the triangle is embedded. In this sort of case, we can say that the pain is gestalt unified with the headache; the experience of the person is gestalt unified with the experience of the rest of the crowd; and the experience of the boundaries is gestalt unified with the experience of the circles.

Whether there are really any cases of gestalt unity is arguable. One could argue that in the above cases, it would be possible to experience the pain without the headache, or have the visual experience of the person without that of the crowd, or to have that of the boundaries without that of the circles, perhaps in some very different context. But it is not implausible that at least some experiences put *some* constraints on concurrent experiences, and that one cannot mix and match experiences arbitrarily. If this is so, then there is at least a weak sort of gestalt unity, since the presence of one phenomenal state puts constraints on the nature of concurrent phenomenal states. In this case one can even say that the presence of one phenomenal state entails the existence of another phenomenal state, where the second is understood as an instantiation of a sufficiently unspecific phenomenal property.

If there is gestalt unity, then there will be cases in which one phenomenal state entails another phenomenal state without the first subsuming the second, at least in any intuitive sense. For example, the experience of the boundary of a Kanisza triangle might entail something about the experience of the nearby objects, but the experience of the nearby objects does not intuitively subsume that of the boundary of the triangle. Similarly, the experience of the shoulder pain might entail the experience of the headache, but it does not intuitively subsume the experience of the headache. (Intuitively, what it is like to have the pain and the headache goes beyond what it is like to have the pain, even if the former is entailed by the latter.)

So: If there is gestalt unity, then subsumption cannot be understood in terms of entailment. But this does not mean that we must give up on the logical unity thesis. Even if subsumption cannot be understood in terms of entailment, one can make a case that the logical unity thesis entails the subsumptive unity thesis.

To see this, we can first note that not *all* phenomenal states are gestalt unified. Even if some pairs of phenomenal states are gestalt unified, it seems very unlikely that all pairs are, and it seems much more plausible that most pairs are not. Given a typical pair of phenomenal states had by a subject such that neither subsumes the other, it usually seems to be straightforwardly possible that a subject could have an instance of the first state without the second. When I see the red book and hear the bird singing, there seems to be no good reason to deny that I could have a visually identical experience without hearing the bird singing, and so on. (Dainton 2000 gives a more extended argument for the conclusion that gestalt unity is not universal and is in fact rare.)

If there can be pairs of states that are not gestalt unified, it also seems that there can be subjects none of whose states are gestalt unified. One simply needs a subject all of whose basic phenomenal states are

independent in the above way: each of them could occur without any of the others. There seems to be no obstacle in principle to such a subject, and one could even argue that our own phenomenal states are often like this. Let us say that such a subject is *gestalt-free*. In *gestalt-free* subjects, the *gestalt* cases of entailment without subsumption will not arise. So (setting aside for a moment any other cases of entailment without subsumption), we can say that if the logical unity thesis holds, the subsumptive unity thesis holds at least when restricted to *gestalt-free* subjects.

Now let us assume that the subsumptive unity thesis holds for *gestalt-free* subjects: for any set of phenomenal states of a *gestalt-free* subject, there is a subsuming phenomenal state. If so, it is very plausible that the subsumptive unity thesis holds for all subjects. If there is always a subsuming state in *gestalt-free* cases, there will plausibly always be a subsuming state in *gestalt* cases. There is nothing about *gestalt* unity that makes the existence of a subsuming state in such cases *less* likely. If anything, the situation is the reverse. In a case of *gestalt* unity, the experiences will be connected in such a way that the existence of a subsuming state will be more likely, not less. So if there are cases in which *gestalt* unified states are not phenomenally unified, there should equally be cases in which *gestalt-free* states are not phenomenally unified. So the subsumptive unity thesis for *gestalt-free* subjects plausibly entails the subsumptive unity thesis for all subjects.

We are close to establishing a connection between the logical unity thesis and the subsumptive unity thesis in general. But we still need to deal with the other case of entailment without subsumption discussed above, in which a subject has distinct simultaneous experiences that share a type. We can deal with this in an analogous way. Let us say that a subject has *duplicate* experiences when the subject has two intuitively distinct experiences that share a maximally specific phenomenal type (two pains or two color experiences with exactly the same quality, say). It is not entirely obvious that duplicate experiences are possible; but in any case, let us say that a *duplicate-free* subject is a subject without duplicate experiences. It is plausible that *if* the subsumptive unity thesis is true when restricted to *duplicate-free* subjects, it is true also of subjects with duplication: if it is possible for duplicate experiences not to be subsumed by a common experience, it will be equally possible for non-duplicate experiences not to be so subsumed. As with *gestalt* phenomena, there is nothing about duplication per se that contributes to a breakdown of phenomenal unity. So the subsumptive unity thesis for *duplication-free* subjects plausibly entails the subsumptive unity thesis for all subjects.

Combining the last two cases, we can say that the subsumptive unity thesis restricted to *gestalt-free*, *duplication-free* subjects plausibly entails the subsumptive unity thesis for all subjects. But it is also clear that the logical unity thesis entails the subsumptive unity thesis for *gestalt-free*, *duplication-free* subjects. In such subjects, a phenomenal state *T* that entails all phenomenal states will also subsume all phenomenal states, since we have removed the relevant gaps between subsumption and entailment. One might worry that one gap remains: by ruling out duplication, we have ruled out the possibility of entailment without subsumption for maximally specific phenomenal states, but two sufficiently nonspecific states of the same type might entail each other without subsuming each other. Nevertheless, since *T* entails maximally specific versions of each of these nonspecific states, *T* will subsume these maximally specific states, and so *T* will subsume the nonspecific states. So *T* subsumes all the subject's phenomenal states.

We have established that the logical unity thesis entails the subsumptive unity thesis for gestalt-free subjects, duplication-free subjects; and we have established that the latter thesis plausibly entails the subsumptive unity thesis for all subjects. So the logical unity thesis plausibly entails the subsumptive unity thesis. In reverse, the subsumptive unity thesis clearly entails the logical unity thesis. So it is plausible that the subsumptive unity thesis holds if and only if the logical unity thesis holds. The only obstacle to this equivalence will arise if there are breakdowns of phenomenal unity that are solely due to gestalt unity or to duplication, but there seems to be little reason to take that possibility seriously.

If this is correct, we can assess the truth of the subsumptive unity thesis by assessing the truth of the logical unity thesis. This latter task is in some respects more straightforward, since we no longer have to deal directly with the primitive notion of subsumption. This also allows the possibility of using familiar logical tools to formulate and assess versions of the unity thesis. We will look more closely at some versions of the thesis in the following section.

6.2 Logical unity and conjunctive closure

There are three versions of the subsumptive unity thesis: the pairwise version, the general version, and the total version. There are correspondingly three versions of the logical unity thesis, holding either that there is logical unity among either any two states of a subject at a time, any set of states, or the complete set of states. Or more directly:

Pairwise Logical Unity Thesis: Necessarily, for any two phenomenal states had by a subject at a time, the subject has a phenomenal state that entails both original states.

General Logical Unity Thesis: Necessarily, for any set of phenomenal states of a subject at a time, the subject has a phenomenal state that entails each state in the set.

Total Logical Unity Thesis: Necessarily, for any conscious subject at a time, the subject has a phenomenal state T such that for any phenomenal state A of the subject at that time, T entails A.

As before, it is clear that the general thesis entails the pairwise thesis and the total thesis as special cases. The total thesis also entails the general thesis and the pairwise thesis, since a state that entails all phenomenal states of a subject will also entail any pair or any set of states. Arguably the pairwise thesis does not entail the other two theses, because of the formal possibility that there might be entailing states for any finite set of states, but not for infinite subsets.

We can start by focusing on the total logical unity thesis, since this corresponds most closely to the total phenomenal unity thesis, which arguably captures the central intuition behind the unity of consciousness. Intuitively, we can think of T, the entailing state in the thesis, as the subject's total phenomenal state, capturing what it is like to be the subject at that time. If such a state exists, it will fulfill the requirement

of the total logical unity thesis.

One can also approach the matter in logical terms. Let us say that the *conjunction* of a set of states is a state C such that necessary, a subject is in C if and only the subject is in each of the states in that set. (Like entailment, conjunction is fundamentally a relation among state-types, and derivatively a relation among state-tokens. Note also that the conjunction of states is quite different from conjunction of the *contents* of states.) This identifies C at least up to mutual entailment. For present purposes, it is useful to assume that when two states A and B mutually entail each other (i.e. when necessarily, a subject is in A if and only if the subject is in B), then the two states are identical. If so, then C is identified uniquely. Nothing that follows rests essentially on this assumption — one could rephrase things in terms of equivalence classes of states — but this makes the discussion easier.

We can then propose a natural candidate for T: the conjunction C of all of a subject's phenomenal states at a time. It is clear that if T exists, T entails C (since T entails each of the conjuncts of C). And it is clear that if T exists, C entails T (since T is itself a phenomenal state). So if T exists, then T is identical to C (and C is therefore a phenomenal state), by the criterion for state identity above. It is also clear that if C is a phenomenal state, then C will satisfy the total logical unity thesis with $T=C$. We can therefore say that an appropriate T exists if and only if C is a phenomenal state.

Let us say that a set of states is *conjunctively unified* when the conjunction of the members of that set is itself a phenomenal state. Then from the discussion above, it follows that the total logical unity thesis is equivalent to the claim that the set of a subject's phenomenal states is conjunctively unified:

Total Conjunctive Unity Thesis: If C is the conjunction of all of a subject's phenomenal states at a time, then C is itself a phenomenal state.

As before, someone might think that a thesis of this sort is trivially true, but this would be incorrect. It is trivial that for any set of phenomenal states of a subject at a time, there will be a conjunctive state C that entails each of the original states. But it is nontrivial that C will itself be a phenomenal state. That is, it is nontrivial (although very plausible) that there will be something it is like to be in C: some global phenomenal character that a subject will have if and only if the subject is in C. Those who deny the original unity thesis will deny the existence of such a phenomenal character, and so will deny that C is itself a phenomenal state.

In effect, we have seen that the original phenomenal unity thesis is equivalent to a thesis about the *conjunctive closure* of co-instantiated phenomenal states (where co-instantiated states are states had by the same subject at the same time): certain conjunctions of states in this class must also be states in this class. This is very useful, since conjunctive closure is amenable to relatively straightforward analysis.

One can also formulate conjunctive closure theses that are closely related to the other versions of the logical unity thesis. There is a pairwise version, and a general version:

Pairwise Conjunctive Unity Thesis: For any two phenomenal states of a subject at a time, their conjunction is a phenomenal state.

General Conjunctive Unity Thesis: For any set of phenomenal states of a subject at a time, their conjunction is a phenomenal state.

These theses are not quite formally equivalent to the corresponding versions of the logical unity thesis. To see this, note that it is at least a formal possibility that two states might be logically unified but not conjunctively unified. For example, it is at least formally possible that the conjunction of *all* of a subject's phenomenal states might be a phenomenal state, but that the conjunctions of certain pairs and subsets might not be. If so, then these pairs and subsets will be logically unified but not conjunctively unified. In this case the pairwise and general conjunctive unity theses will be false, but the the pairwise and general logical unity theses will be true.

However, it is clear that these conjunctive unity theses *entail* the corresponding versions of the logical unity thesis. And they are interesting and plausible theses in their own right. The first says that for any two phenomenal states A and B of a subject at a time, there will be something distinctive it is like to be in A and B: that is, a distinctive conjoint phenomenal character that a subject will have if and only if the subject is in both A and B. The second says the same thing for arbitrary sets of co-instantiated phenomenal states. These theses are not formally trivial, but they are highly plausible theses about phenomenal consciousness. (These theses are closely related to the correspondence thesis discussed in the previous section.)

All three theses are simple and elegant. The pairwise conjunctive unity thesis says that the class of phenomenal states is closed under pairwise co-instantiated conjunction: the conjunction of two co-instantiated phenomenal states is a phenomenal state. The general conjunctive unity thesis says that the class of phenomenal states is closed under general co-instantiated conjunction: the conjunction of any set of co-instantiated phenomenal states is a phenomenal state. And the total conjunctive unity thesis says that the class of phenomenal states is closed under maximal co-instantiated conjunction: the conjunction of a maximal set of co-instantiated phenomenal states is a phenomenal state.

The total conjunctive unity thesis remains the core version of the unity thesis, but all of these theses are plausible and useful. Each of them can be used as a tool in assessing the status of the unity of consciousness, in assessing its consequences, and in assessing its compatibility with various theories of consciousness.

6.3 Hurley, Shoemaker, and what it is like

It should be noted all of these unity theses are stated simply in terms of the notions of phenomenal state, of co-instantiation, and of conjunction. And the notion of a phenomenal state is tied constitutively to the notion of there being something it is like to be a given subject, or to be in a given state. So we have an account of unity that requires little more than the existing "what it is like" conception of phenomenal

states.

This stands in tension with a claim in a very interesting analysis by Hurley (1998; this volume). Hurley (1998, pp. 165-66) argues that the unity of consciousness cannot be characterized "subjectively", and that suppositions about the structure of consciousness are not captured by the "what it is like" test, so that we need to appeal to further "objective" properties to give an account of unity. This claim is grounded in the claim that in a case where unity breaks down, there is no "what-it-is-like" that captures the structure of a subject's consciousness. Hurley backs up this claim by considering two cases: (i) two subjects, one experiencing red and hot, the other experiencing red and dizzy; and (ii) a partially unified single subject, in whom red and hot are unified, red and dizzy are unified, but hot and dizzy are not. Hurley argues that no "what it is like" facts can distinguish these two cases.

But from the claim that there is no what-it-is-like that characterizes a disunified subject, it does not follow that one cannot characterize unity in what-it-is-like terms. Indeed, following Hurley's own claim, one can hold that unity breaks down precisely when there is nothing it is like to have all a subject's conscious states simultaneously. We can distinguish case (i) from case (ii) above by noting that in case (ii), both subjects have a phenomenal state that subsumes all their phenomenal states, whereas in case (i), the subject has no such phenomenal state. Of course, our characterization of unity appeals to something more than phenomenal states themselves: it appeals to subsumption, and to co-instantiation in a subject. Perhaps Hurley would count these notions as in some sense "objective". There is no point arguing over terminology here, but we can at least note that subsumption is a phenomenal relation, fixed by phenomenology alone: if A subsumes B, then the phenomenology of A guarantees that it subsumes B. And subjects are simply the bearers of phenomenal states. So we are staying quite close to home in characterizing unity this way.

Hurley might extend her argument by suggesting case (iii): a bifurcated subject with two different (but indistinguishable) tokens of red in separate streams. In this subject, red1 is unified with hot, red2 is unified with dizzy, and no state in either pair is unified with a state in the other pair (Hurley 1998, p. 166, seems to point toward such a case). If (iii) is possible, one could argue that it could not be distinguished from (ii) by talk of subjects and their phenomenal states alone. We would need to appeal to the identity of phenomenal states: a single "red" experience is involved in both complex experiences in (ii), but not in (iii).

There are a number of things one could say in response. One might concede that "what it is like" talk cannot distinguish the two different cases of disunity (ii) and (iii), but hold that it can nevertheless distinguish unity from disunity, which is the most important work we need it to do. If the unity thesis is true, then cases of disunity will be impossible, and distinctions among impossible cases will not matter to characterizing the structure of consciousness. More deeply, one can suggest that Hurley's argument shows at best that one cannot distinguish the cases in terms of the distribution of phenomenal state-*types*. If we appeal to facts about the distribution of phenomenal state-*tokens*, things are straightforward: there is a token experience that is subsumed by two different complex experiences in (ii), but not in (iii). It may be that (ii) and (iii) will be introspectively indistinguishable, so that the structure of consciousness is not *transparent* to a subject. But nevertheless, a characterization of the structure of consciousness in terms of

phenomenal relations among phenomenal state-tokens is still, in a deep sense, a characterization in subjective terms.

Our characterization of unity in phenomenal terms also stands in tension with a claim by Shoemaker (this volume). Shoemaker suggests that if a conscious state is understood as one with a phenomenal property (i.e. one such that there is something it is like to be in it), this leads to "consciousness atomism": the view that the factors that make a state conscious are independent of the factors that make two states unified. Our discussion here suggests that this is false. What it is for two conscious states to be unified can be understood in terms of the existence of a more complex conscious state, where both the simple state and the complex states are states characterized by what it is like to be in them. So the factors that enter into unifying conscious states are the same sort of factors as those that enter into those states being conscious in the first place.

At one point, Shoemaker characterizes "consciousness atomism" differently, as the view that "whether a state is conscious will be independent of whether there are other conscious states with which it is co-conscious". The account here is neutral on this claim. For all we have said here, it may be possible for there to be a subject with a single conscious state. This claim does not seem to us to be obviously objectionable, and it is compatible with the more important view that the factors that enter into consciousness are the same as those that enter into co-consciousness.

In fact, the definitions of unity that we have given here suggest that any account of what it is to be a phenomenal state will automatically yield a theory of what it is for two such states to be unified. We need simply to apply the theory to the relevant conjunctive states, in order to determine whether they are phenomenal states. In this way, it seems that any substantive theory of phenomenal consciousness can yield unified definitions of consciousness and of co-consciousness. It is precisely because of this that the unity thesis (if it is true) puts strong constraints on a theory of phenomenal consciousness, as we will see.

7 Applications of the Unity Thesis

We have already mentioned the objection that the conjunctive versions of the unity thesis are trivial: that is, that it is trivial that the conjunction of a set of co-instantiated phenomenal states is itself a phenomenal state. It is clear that the thesis is not *formally* trivial, in that there are many classes of states that are not closed under co-instantiated conjunction: e.g., states of the sort "talking with X", where X is an individual. Closer to home: there are also many classes of *mental* states that are not closed under co-instantiated conjunction.

For example, the class of *belief* states does not seem to be closed under conjunction. Let us say that a belief state is the state of believing some proposition. Then it is not the case that the conjunction of any set of belief states is a belief state. For example, if A is the state of believing that P, and B is the state of believing that Q, there is plausibly no belief state that a subject will be in precisely when they are in A and B. The only tenable candidate for such a belief state is the state of believing P and Q. But there are well known reasons to believe that a subject can believe that P and believe that Q without believing the

conjunction $P \wedge Q$. For example, P and Q might be believed in different "compartments" of a compartmentalized mind. It may even be that for some P, a subject can believe that P and separately believe that $\sim P$, without believing the contradiction $P \wedge \sim P$. And it seems quite possible that a subject can have many different beliefs without accepting the massive conjunction of the contents of all of those beliefs. If this is right, then the conjunction of co-instantiated belief states will not in general be a belief state. So the class of belief states is not closed under co-instantiated conjunction.

It may seem plausible or even obvious that the class of *phenomenal* states is closed under conjunction. But if so, this is a substantive thesis about the class of phenomenal states, and their difference with other classes of mental states. It may even be a *conceptual* truth, in some sense, that the class of phenomenal states is closed under co-instantiated conjunction. But if so, this is again a substantive thesis about the *concept* of a phenomenal state, and a way in which it differs from the concept of a belief state, and of other sorts of states.

The substantive nature of the thesis is revealed by the fact that the conjunctive unity theses put a strong constraint on potential theories of consciousness. We have seen that the unity thesis is *prima facie* plausible, and there there seem to be no strong arguments against it. If this is right, then the unity thesis puts a *prima facie* constraint on theories of consciousness: they must be compatible with the unity thesis. And in particular, any account of phenomenal states must be compatible with the total conjunctive unity thesis. Whatever phenomenal states are, according to a given account, the class of phenomenal states must be closed under total co-instantiated conjunction. A number of prominent theories of consciousness appear to be incompatible with this constraint.

One example is the higher-order thought theory of consciousness, put forward by David Rosenthal (1997) and others. Not all higher-order thought theorists intend the theory as an account of phenomenal consciousness (e.g. Lycan 2000 explicitly rejects the idea), but we are only concerned with versions of the theory that are aimed at phenomenal consciousness. The central idea of these theories is the following:

Higher-Order Thought Thesis: A mental state M is phenomenally conscious if and only if a subject has a higher-order thought about M.

Here, a higher-order thought about M should be understood as a thought by the subject with the content "I am in M". The thesis will usually be modified and qualified in some ways. For example, Rosenthal holds that for M to be conscious, the higher-order thought must be brought about in the right sort of way, and in particular must be a noninferential thought. Rosenthal also holds that only sensory states can be phenomenally conscious, so that we would have to insert a rider to that effect in the definition above. This is arguably a mere terminological difference, however, since Rosenthal holds that there will be something it is like to be in a state whenever it is the object of the right sort of higher-order thought, whether the state is sensory or not. In any case, for our purposes we will take the thesis in the simple form above. Our arguments should apply straightforwardly to most modified versions.

Is the higher-order thought thesis compatible with the unity thesis? It is easiest to approach this question by considering the conjunctive versions of the unity thesis. The conjunctive versions say that the class of phenomenal states are closed under conjunction. So we can ask: on the higher-order thought theory, is the class of phenomenal states closed under conjunction?

We can start by thinking about phenomenally conscious mental states. If A and B are phenomenally conscious mental states, is A&B necessarily a phenomenally conscious mental state? Assuming the higher-order thought thesis, this translates into the following: if a subject has a higher-order thought about A and a higher-order thought about B, does the subject necessarily have a higher-order thought about A&B? That is, if the subject has a thought "I am in A" and a thought "I am in B", does it follow necessarily that the subject has a thought "I am in A and B"?

It seems not. It is surely possible for a subject to think "I am in A" and "I am in B", without connecting these into a thought "I am in A and B". We can take a case like those discussed above, in which a subject has contradictory beliefs, knows that she has each belief, but never puts the two together. She might have the thought "I believe P" and the thought "I believe ~P" without ever putting these two together into a thought "I believe both P and ~P". This might be strange or unusual, but there is nothing contradictory about it. There would only be something contradictory here if the beliefs of a subject are necessarily closed under logical consequence; but of course no subject's beliefs are closed under logical consequence.

The same is even clearer where total conjunctivity is concerned. On the higher-order thought theory, if a subject has a number of phenomenally conscious mental states, is their conjunction a phenomenally conscious mental state? That is, if a subject has mental states $A_1, \dots, A_n$, and has the thoughts "I am in A_1 ", ..., "I am in A_n ", does the subject necessarily have the thought "I am in $A_1 \& A_2 \& \dots A_n$ "? Again, it seems not. One might reasonably argue that this entailment does not even hold *typically*, let alone necessarily. That is, it is arguable a *typical* subject with these higher-order beliefs would not have the complex conjunctive belief. Whatever one says here, it is hard to dispute that it is *possible* for a subject to have the individual higher-order beliefs without the complex conjunctive belief.

So it appears that if the higher-order thought view is true, the class of phenomenally conscious mental states is not closed under co-instantiated conjunction. This already contradicts the central intuition behind the unity thesis: that necessarily, if there is something it is like to be in each of a set of states, there is something it is like to be in all the states at once. On the higher-order thought view, this thesis will clearly be false.

The official version of the unity thesis is stated in terms of phenomenal states, not phenomenally conscious mental states. The analysis of phenomenal states is slightly trickier, since advocates of the higher-order thought view have not usually talked about phenomenal states and phenomenal properties directly. But given that higher-order thought theorists hold that there is something it is like to be in a mental state when the subject has a higher-order thought about it, they presumably hold that what it is like to be in that state is determined by the content of the higher-order thought. If so, it seems that phenomenal properties will be the properties of having higher-order thoughts with certain contents, and

phenomenal states will be the states of having such higher-order thoughts.

Do phenomenal states, understood this way, satisfy the unity thesis? It seems not, for much the same reason as before. Here it is useful to take the entailment version of the unity thesis: that necessarily, when a subject has a set of phenomenal states, the subject has a phenomenal state that entails each of the individual states. When a subject has a set of higher-order thoughts $H_1, \dots, H_n$, does the subject necessarily have a higher-order thought HH such that being in HH entails being in $H_1, \dots, H_n$? It seems not, for the usual reasons. A subject might think "I am in A" and "I am in B", without any higher-order thought (e.g. "I am in A&B") such that having that thought entails having the original thoughts.

The problem is not that the higher-order thought theory provides no way to understand phenomenal unity. It can do so in a natural way. Two phenomenally conscious mental states A and B are unified when the subject has a higher-order thought about them not just singly but jointly. And two phenomenal states, the states of having higher-order thoughts "I am in A" and "I am in B", are phenomenally unified when there is a complex phenomenal state that entails them: that is, if there is a complex higher-order thought such that having the complex thought entails having the specific thoughts. This requirement will arguably be satisfied when the subject has a complex higher-order thought such as "I am in A&B".

The problem is rather that on this account, there is no reason to believe that phenomenal states, or phenomenally conscious mental states, will always be unified. Certainly it will not be necessary that they be unified, and it seems plausible that in a typical case, they will not be unified. So the higher-order thought thesis is incompatible with the unity thesis. It is clearly incompatible with the conjunctive and logical versions of the unity thesis. It is therefore also incompatible with the subsumptive versions, since any failure of logical unity automatically entails a failure of subsumptive unity. So if the higher-order thought thesis is true, the unity thesis is false. And if the unity thesis is true, the higher-order thought thesis is false.

Proponents of the higher-order thought thesis might reply in a number of ways. Most straightforwardly, they might reply by denying the unity thesis. This is a tenable response, since the truth of the unity thesis cannot be taken for granted. But still, there is a strong intuition that the unity thesis is true, so the incompatibility is at least a cost of the higher-order thought thesis. Proponents might also embrace a more limited version of the unity thesis, arguing for example that unity holds typically but not necessarily, or that it holds given contingent facts of human psychology, but not for all possible beings. Here there would still be the cost of denying the intuition of necessary unity, and there would be the added difficulty of defending the claim that unity holds in the relevant range of cases, when there seems to be no obvious reason why complex conjunctive thoughts about all the objects of our higher-order thoughts should typically exist.

A higher-order thought theorist might also respond by finding fault with the argument for incompatibility: they might hold, for example, that it is necessary that the class of mental states that are objects of higher-order thoughts is closed under conjunction. This would be a difficult case to make, in face of the apparent possibility of failure of this principle, and in face of the general phenomenon that

beliefs are not closed under logical consequence.

Finally, a proponent might modify the higher-order thought thesis to make it compatible with the unity thesis. To do so, they must modify the definition of a phenomenally conscious mental states. It could be held, for example, that a mental state is phenomenally conscious when either (i) it is the object of a higher-order thought, or (ii) it is the conjunction of states that are the objects of higher-order thoughts. This sort of disjunctive account would be contrary to both the letter and the spirit of existing higher-order thought views (which hold that a conscious state is one that the subject is conscious of). One could also raise questions about whether this thesis delivers any *substantive* unity of consciousness, or merely a stipulated sort of unity of consciousness that holds trivially. And so far as the unity of consciousness seems to be a substantive fact about consciousness, one could argue that this modified version of the higher-order thesis does not really account for it.

Of course all of this is debatable and could lead to fruitful further discussion. But the *prima facie* incompatibility between the two theses is at least interesting. It is worth noting that the incompatibility extends straightforwardly to other "higher-order" views of consciousness, including views on which a conscious state is an object of a higher-order perceptual state, or the object of some other sort of higher-order representational state. The existence of a set of higher-order perceptual states does not entail the existence of a complex conjunctive higher-order perceptual state, and the same goes for other sorts of representational states. So if the unity thesis is true, these theses are false, and vice versa.

The unity thesis is also incompatible with many *representationalist* views of consciousness. According to representationalist views (e.g. Dretske 1995, Tye 1995), all phenomenally conscious mental states are representational states (that is, states with representational content). This is commonly allied with a further functional criterion to yield:

Representationalist Thesis: a mental state is phenomenally conscious if and only if it is a representational state that plays an appropriate functional role.

We will focus on this broadly functionalist variety of representationalism. The details of the relevant functional role differ between representationalists, but it is typically held to involve some sort of access and control. One can then say that what it is like to be in a mental state is determined by the content of the representational state, on the condition that it plays the relevant functional role. On this sort of view, then, a phenomenal state is a state of having a certain sort of representational state play the appropriate functional role, where distinct phenomenal states are individuated by distinct representational contents.

Two phenomenal states P1 and P2 are conjunctively unified when there is a phenomenal state P that entails each of the original states. On the representationalist account, two phenomenal states P1 and P2, corresponding to representational states A1 and A2 (with contents C1 and C2) playing the relevant functional role, will be conjunctively unified when there is a phenomenal state P, corresponding to representational state A (with content C) playing the relevant functional role, such that P entails P1 and

P2. This will occur if and only if the existence of A playing the role entails the existence of both A1 and A2 playing the role. The only reasonable way to satisfy this is for the content of A to entail the content of A1 and the content of A2: that is, for C to entail both C1 and C2, or for C to entail C1&C2. For example, if A1 has content "red to the left", and A2 has content "green to the right", P1 and P2 will be conjunctively unified if there is a state A (playing the role) whose content entails "red to the left and green to the right". So two phenomenal states, corresponding to two representational states, will be conjunctively unified if and only if there is a conjunctive representational state (playing the appropriate role) whose content entails the conjunction of the contents of the original representational states.

The unity thesis is true if and only if necessarily, every set of phenomenal states is conjunctively unified. On the representationalist view, is this the case? It seems not. It seems at least possible to have a state with content C1 and a state with content C2, each playing a certain role, without having a state with content C1&C2 that plays the role. We saw this earlier in the case where the relevant role involves accessibility: it is possible that C1 is accessible and C2 is accessible with C1&C2 being inaccessible. Something similar will hold for any functional role involving access and control. If this is so, then representationalist thesis in the relevant class are incompatible with the unity thesis.

As before, representationalists could respond in a number of ways. They could deny the unity thesis, at cost of denying a strong intuition. They could modify it to apply to a more limited range of cases, at cost of some intuition and perhaps some empirical constraint. (For example, in the Sperling case, this representationalist may have to deny that the subject has a phenomenally unified visual field.) They could modify the representationalist thesis to allow a disjunctive definition on which is stipulated that conjunctions of phenomenal states are phenomenal states, at cost of endangering the substantive status of the unity thesis. Alternatively, they could move to a different sort of representationalism which is not so closely tied to functionalism: for example, it might be held that phenomenally conscious states are representational states whose content is represented *phenomenally*, or that they are representational states with some other property that is not functionally defined. The resulting version of representationalism might be compatible with the unity thesis (as well as being independently more plausible than the previous versions), at cost of giving up the reductive aspirations of many representationalist views.

One might also argue that other nonrepresentationalist forms of functionalism are incompatible with the unity thesis, on the general grounds that there will not be the relevant conjunctive property among states playing the functional role. The details will depend on the details of the functionalist theory, and in particular on the account that is given of phenomenal states and properties. These accounts can vary between functionalist theories, and are often not clearly articulated, so it is difficult to give a general analysis of such theories with respect to the unity thesis. But it is clear that it will be at least highly nontrivial for a functionalist account to satisfy the unity thesis.

If what has gone above is correct, then the unity thesis is incompatible with higher-order thought (and other higher-order representation) views of consciousness, with many representationalist views of consciousness, and with many functionalist views of consciousness. So the unity thesis is clearly nontrivial. Nevertheless, it has strong independent plausibility as a thesis about phenomenal states. So the incompatibility of the unity thesis with these views of consciousness should be seen as at least a *prima*

facie argument against these views.

8 Explaining the Unity Thesis

If the unity thesis is true, how is its truth to be explained? We do not know the answer to this question. But in this concluding section, we will explore some possibilities.

One common strategy is to try to explain unity in functional terms. For example, one might try to explain unity in terms of some sort of informational integration, or in terms of serial processing in the brain, or something along those lines. One obvious problem with this sort of strategy is that it is not clear why this sort of functioning should yield phenomenal unity, as opposed to something like access unity. But an equally deep problem is that for reasons similar to those discussed above, it seems inevitable that this sort of functioning will be present *contingently*, and that it will be possible for conscious states to exist that do not stand in the relevant functional relations. If so, unity (on these analyses) will obtain only contingently, and the unity thesis will be false. If unity is to obtain necessarily, as the unity thesis suggests, we must look elsewhere.

Much of the reason for accepting the truth of the unity thesis comes from the fact that its denial seems to be inconceivable, and perhaps incoherent. This suggests that the unity thesis may be at some level a *conceptual* truth, although perhaps a deep conceptual truth, whose roots are revealed only by a deep analysis of our concepts. The central concepts involved in the unity thesis are that of a phenomenal state and that of a subject, along with various additional notions such as subsumption, entailment, conjunction, and so on. So one might hope that some light could be shed by attention to the concept of a subject, or by attention to the concept of consciousness.

One natural suggestion is that our concept of a subject of experience is somehow premised on unity. For example, one could suggest that ascriptions of subject-hood require as a precondition that subjects correspond to unified phenomenal fields. In the spirit of a sort of bundle theory of the subject, one could argue that we have a prior notion of a phenomenal field, and that we then associate subjects with phenomenal fields. If this is the case, we would expect that every subject would have a unified consciousness. A subject with two distinct phenomenal fields, for example, would be ruled out as a conceptual impossibility: where there are two phenomenal fields, there will automatically be two subjects.

How might this work? Our articulation of the notion of a phenomenal field in this paper appeals to subjects and co-instantiation, but one might argue that these can be bypassed. For example, one might appeal to a primitive relation of subsumption (or of co-consciousness) among phenomenal states that makes no presuppositions about subjects of those states, and then define a phenomenal field as a maximal phenomenal state: a phenomenal state that is not subsumed by any other phenomenal state. But even if something like this works, there is a deeper problem. This strategy might explain why distinct phenomenal fields correspond to distinct subjects, but it cannot explain why states of consciousness come packaged into unified phenomenal fields in the first place. For example, nothing in this strategy explains

why a phenomenal state cannot be subsumed by two different phenomenal states such that no further phenomenal state subsumes both of these in turn. More generally, nothing here explains why the subsumption relation does not hold in quite unsystematic and fragmented manner. It is possible that an analysis of subsumption itself could do some work: for example, one could argue that subsumption is conceptually akin to a mereological part-whole relation, and so must hold reflexively, antisymmetrically, and transitively, and perhaps in a way that allows no overlap. But this conceptual stipulation does not really make the problem go away. It simply raises the question of why conscious states come packaged as parts and wholes.

One might then take a different approach. Instead of focusing on the concept of a subject, one can focus on the concept of consciousness itself. It could be argued that our *basic* concept of consciousness is not the notion of a simple phenomenal state — what it is like to such-and-such at a time. Rather, our basic notion of consciousness is that of a total phenomenal state: what it is like to be a subject at a time. This yields a holistic rather than an atomistic view of consciousness. On this approach, we do not start with basic atomic states of consciousness, and somehow glue them together into complex states. Rather, we start with a basic *total* state of consciousness, and then differentiate it into simpler states, and ultimately into atomic states.

If this were truly our basic notion of consciousness, then it might explain why the unity thesis is true. On this view, any non-total phenomenal state is *derivative* on a total phenomenal state that subsumes it. On this view, it is to be expected that any phenomenal states of a subject at a time are all simply aspects of what it is like to be that subject at that time. As such, it is to be expected that for any set of co-instantiated phenomenal states, there will be a subsuming state. On this view, the most basic problem with the theories of consciousness discussed in the last section is that they are atomistic rather than holistic, starting with simple states rather than total states. If this view is right, then any such analysis of consciousness will be a misanalysis from the start.

It is not obvious that this sort of conceptual claim on its own yields a substantive unity thesis. But one might naturally tie this analysis to a corresponding view of the metaphysics of consciousness. In nature, it may be that the most basic sort of conscious state is the total phenomenal state, or the phenomenal field, or even the phenomenal world. These total states are basic, but they are not featureless: they come with a complex structure from which one can differentiate many aspects. (As an analogy, one can think of a quantum wavefunction, which is a basic state in physics but which nevertheless has a complex structure.) So metaphysically, simple conscious states might be derivative on total conscious states. If so, we would have a clean explanation of why a substantive unity thesis is true.

This sort of suggestion is highly speculative, and much needs to be worked out. For example, it is far from obvious that our basic concept of consciousness is that of a total state of consciousness, and one needs to make a direct case for this. And the corresponding metaphysics needs to be worked out in much more depth. But there is at least some plausibility in the idea that the concept of consciousness, and states of consciousness, are fundamentally holistic rather than atomistic. And this squares well with our intuition that consciousness is necessarily unified.

In any case, whether the substantive claims that we have made in this paper are correct or incorrect, we hope to have helped to pin down some of the crucial issues. It is clear that there is much need for further work in analyzing the notion of unity, in assessing the truth of the unity thesis, and in seeking an explanation of its truth. It is likely that such work will be philosophically fruitful.

Acknowledgments

Versions of this paper have been presented at the ASSC conference on the unity of consciousness in Brussels, at a Pew conference on the metaphysics of human beings in Syracuse, at a conference on neurophenomenology in Boulder, and in talks at Alabama, Arizona, Barcelona, University College London, and Wake Forest. We thank audience members on all those occasions for helpful comments. Special thanks to Barry Dainton, Bernie Kobes, Graham Macdonald, and Paul Studtmann.

References

- Bayne, T. 2001. Co-consciousness: Review of Barry Dainton's *Stream of Consciousness*. *Journal of Consciousness Studies* 8:79-92.
- Block, N. 1995. On a confusion about a function of consciousness. *Behaviora and Brain Sciences* 18:227-47.
- Dainton, B. 2000. *Stream of Consciousness: Unity and continuity in conscious experience*. Routledge.
- Dennett, D.C. 1992. The self as a center of narrative gravity. In (F. Kessel, P. Cole, & D. Johnson, eds), *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Erlbaum, 1992.
- Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.
- Hurley, S. 1998. Unity, neuropsychology, and action. In *Consciousness in Action*. Harvard University Press.
- Hurley, S. 2002. Action, the unity of consciousness, and vehicle externalism. In (A. Cleeremans, ed) *The Unity of Consciousness: Binding, Integration, Dissociation*. Oxford University Press.
- Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Blackwell.
- Lycan, W. 2000. Representational theories of consciousness. In (E. Zalta, ed) *The Stanford Encyclopedia of Philosophy*.
- Marcel, A.J. 1993. Slippage in the unity of consciousness.

Marcel, A.J. 1994. What is relevant to the unity of consciousness? In (C. Peacocke, ed) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford University Press.

Marks, C. 1980. *Commissurotomy, Consciousness, and Unity of Mind*. MIT Press.

Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22:396-413. Reprinted in *Mortal Questions* (Cambridge University Press, 1979).

Puccetti, R. 1981. The case for mental duality: Evidence from split-brain data and other considerations. *Behavioral and Brain Sciences* 4:93-123.

Revonsuo, A. 1999. Binding and the phenomenal unity of consciousness, *Consciousness and Cognition* 8: 173-185.

Rosenthal, D.M. 1997. A theory of consciousness. In (N. Block, O. Flanagan, and G. Güzeldere, eds) *The Nature of Consciousness*. MIT Press.

Shoemaker, S. 2002. Consciousness and co-consciousness. In (A. Cleeremans, ed) *The Unity of Consciousness: Binding, Integration, Dissociation*. Oxford University Press.

Sperling, G. 1960. The information available in brief visual presentations. *Psychological Monographs* 498:1-29.

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press.

Contemporary Philosophy of Mind: An Annotated Bibliography

Compiled by [David Chalmers](#), Department of Philosophy, University of Arizona, Tucson, AZ 85721. E-mail: chalmers@arizona.edu. (c) 2001 David J. Chalmers. Last updated: July 15, 2001.

This is a bibliography of recent work in the philosophy of mind, philosophy of cognitive science, philosophy of artificial intelligence, and on consciousness in the sciences. It consists of 5702 entries, and is divided into six parts, each of which is further divided by topic and subtopic.

- [Part 1: Consciousness and Qualia](#) [1082 entries]
- [Part 2: Mental Content](#) [869 entries]
- [Part 3: Metaphysics of Mind](#) [1026 entries]
- [Part 4: Philosophy of Artificial Intelligence](#) [565 entries]
- [Part 5: Philosophy of Psychology](#) [561 entries]
- [Part 6: Consciousness in the Sciences](#) [1599 entries]

Many of the entries are annotated with a brief summary. (I don't necessarily take responsibility for these annotations, many of which I wrote in the distant past.) The bibliography consists mostly of papers and books from the last few decades, with density of coverage increasing with proximity to the present. Most of the core areas of contemporary philosophy of mind are covered, but some areas (e.g. (philosophy of perception, philosophy of action, propositional attitude semantics, moral psychology) receive less coverage than others. A fuller table of contents is below.

Part 6 is relatively new and is not annotated. It's not strictly "philosophy of mind", but I've put it here for convenience. Of course the division between parts is inexact. There is plenty of material by philosophers in part 6, and plenty of material by scientists in the other parts (especially parts 1, 4, and 5).

The six parts here are in HTML. You can also get ASCII versions of [Part 1](#), [Part 2](#), [Part 3](#), [Part 4](#), [Part 5](#), [Part 6](#), and the [whole thing](#). These can be useful for systematic use when downloaded, in combination with standard search functions.

I can no longer spend as much time on this bibliography as I used to, but I try to make a fairly thorough update once a year. New entries are rarely annotated, however. The online versions of *Current Contents*, *PsycInfo*, and the *Philosophers' Index* are all invaluable in this process. Thanks to Brad Thompson for his help with the 1999 and 2001 updates. I am grateful for pointers to any relevant material that I have missed.

For links to other bibliographies, and for links to around 1000 online papers in many of these areas, see

my

- [\(other\) online bibliographies on philosophy, consciousness, and such](#)
- [online papers on consciousness](#)

Comments, corrections, and suggestions for additions are more than welcome.

Part 1: Consciousness and Qualia [1082]

- 1.1 [Consciousness](#) [85]
 - 1.1a [General](#) [47]
 - 1.1b [The Concept of Consciousness](#) [38]
- 1.2 [Explaining Consciousness?](#) [115]
 - 1.2a [Subjectivity and Objectivity](#) (Nagel) [46]
 - 1.2b [The Explanatory Gap](#) (Levine) [16]
 - 1.2c [`Hard' and `Easy' Problems](#) (Chalmers) [25]
 - 1.2d [Cognitive Closure](#) (McGinn) [16]
 - 1.2e [Miscellaneous](#) [12]
- 1.3 [Materialism and Dualism](#) [213]
 - 1.3a [The Knowledge Argument](#) (Jackson) [46]
 - 1.3b [Zombies & Modal Arguments](#) [38]
 - 1.3c [Essentialist Arguments](#) (Kripke) [31]
 - 1.3d [Arguments from Disembodiment](#) [17]
 - 1.3e [Consciousness and Physicalism, Misc](#) [14]
 - 1.3f [Consciousness and Dualism](#) [16]
 - 1.3g [Mind-Body Problem, Misc](#) [51]
- 1.4 [Specific Views on Consciousness](#) [227]
 - 1.4a [Higher-Order Thought Approaches](#) (Rosenthal, etc) [46]
 - 1.4b [Dennett on Consciousness](#) [60]
 - 1.4c [Functionalism](#) [8]
 - 1.4d [Eliminativism](#) [16]
 - 1.4e [Epiphenomenalism](#) [30]
 - 1.4f [Interactionism](#) [16]
 - 1.4g [Panpsychism](#) [24]
 - 1.4h [Intrinsic Monism](#) (Russell, Feigl, etc) [27]
- 1.5 [Consciousness and Content](#) [129]
 - 1.5a [Consciousness and Intentionality](#) (Searle, etc) [23]

- 1.5b [The Content of Experience](#) [48]
 - 1.5c [Representationalism](#) [38]
 - 1.5d [Internalism and Externalism about Experience](#) [12]
 - 1.5e [Miscellaneous](#) [8]
- 1.6 [Aspects of Consciousness](#) [182]
 - 1.6a [Self-Consciousness](#) [54]
 - 1.6b [Unity of Consciousness](#) [45]
 - 1.6c [Homogeneity of Consciousness](#) (Sellars, etc) [11]
 - 1.6d [Knowledge of Consciousness](#) [50]
 - 1.6e [The Function of Consciousness](#) [22]
- 1.7 [Qualia](#) [131]
 - 1.7a [Qualia, General](#) [27]
 - 1.7b [Qualia and Materialism](#) [13]
 - 1.7c [Eliminativism about Qualia](#) [11]
 - 1.7d [The Inverted Spectrum](#) [31]
 - 1.7e [Absent Qualia](#) (Block, etc) [15]
 - 1.7f [Introspection and Absent Qualia](#) (Shoemaker) [10]
 - 1.7g [Functionalism and Qualia, General](#) [24]

Part 2: Mental Content [869]

- 2.1 [Propositional Attitudes and Folk Psychology](#) [285]
 - 2.1a [The Language of Thought](#) (Fodor) [54]
 - 2.1b [The Intentional Stance](#) (Dennett) [40]
 - 2.1c [Eliminativism](#) (Churchlands, etc) [57]
 - 2.1d [Propositional Attitudes, General](#) [42]
 - 2.1e [The Nature of Folk Psychology](#) [38]
 - 2.1f [The Simulation Theory](#) [54]
- 2.2 [Internalism and Externalism](#) [278]
 - 2.2a [Is Content in the Head?](#) (Putnam, Burge) [38]
 - 2.2b [Externalism and Psychological Explanation](#) [41]
 - 2.2c [Externalism and Mental Causation](#) [28]
 - 2.2d [Externalism and the Theory of Vision](#) [16]
 - 2.2e [Externalism and Computation](#) [11]
 - 2.2f [Externalism and Self-Knowledge](#) [72]
 - 2.2g [The Status of Narrow Content](#) [43]
 - 2.2h [Miscellaneous](#) [29]
- 2.3 [Theories of Content](#) [144]

- 2.3a [Information-Based Accounts](#) (Dretske, etc) [29]
- 2.3b [Asymmetric Dependence](#) (Fodor) [20]
- 2.3c [Causal Accounts, General](#) [17]
- 2.3d [Teleological Approaches](#) (Millikan, etc) [48]
- 2.3e [Conceptual Role Approaches](#) [19]
- 2.3f [Theories of Content, Misc](#) [11]
- 2.4 [Representation, General](#) [25]
- 2.5 [The Explanatory Role of Content](#) (Dretske, etc) [22]
- 2.6 [Concepts](#) [46]
- 2.7 [Meaning Holism](#) [26]
- 2.8 [Mental Content, Misc](#) [43]

Part 3: Metaphysics of Mind [1026]

- 3.1 [Supervenience](#) [94]
 - 3.1a [Psychophysical Supervenience](#) (Kim, etc) [14]
 - 3.1b [Supervenience and Physicalism](#) [25]
 - 3.1c [Technical Issues in Supervenience](#) [27]
 - 3.1d [Supervenience, General](#) [28]
- 3.2 [Reduction](#) [82]
 - 3.2a [Multiple Realizability and Reduction](#) [25]
 - 3.2b [Nonreductive Materialism](#) [28]
 - 3.2c [Reduction in Psychology](#) [14]
 - 3.2d [Miscellaneous](#) [15]
- 3.3 [Other Psychophysical Relations](#) [127]
 - 3.3a [Physicalism](#) [29]
 - 3.3b [Token Identity](#) [7]
 - 3.3c [Emergence](#) [58]
 - 3.3d [Dualism](#) [21]
 - 3.3e [Miscellaneous](#) [12]
- 3.4 [Functionalism](#) [63]
 - 3.4a [Causal Role Functionalism](#) (Armstrong/Lewis) [17]
 - 3.4b [Machine Functionalism](#) (Putnam) [15]
 - 3.4c [Miscellaneous](#) [31]
- 3.5 [Other Psychophysical Theories](#) [238]
 - 3.5a [Logical Behaviorism](#) (Ryle, etc) [53]
 - 3.5b [Identity Theory](#) (Smart, Place, etc) [98]
 - 3.5c [Eliminative Materialism](#) (Rorty, Feyerabend) [30]

- 3.5d [Anomalous Monism](#) (Davidson) [57]
- 3.6 [Mental Causation](#) [58]
- 3.7 [Personal Identity](#) [189]
 - 3.7a [Personal Identity, General](#) [76]
 - 3.7b [Parfit on Personal Identity](#) [49]
 - 3.7c [Persons](#) [23]
 - 3.7d [Split Brains](#) [22]
 - 3.7e [Multiple Personality Disorder](#) [19]
- 3.8 [Free Will](#) [129]
- 3.9 [The Problem of Other Minds](#) [46]

Part 4: Philosophy of Artificial Intelligence [565]

- 4.1 [Can Machines Think?](#) [224]
 - 4.1a [The Turing Test](#) [51]
 - 4.1b [Godelian Arguments](#) (Lucas, Penrose) [59]
 - 4.1c [The Chinese Room](#) (Searle) [58]
 - 4.1d [Machine Consciousness](#) [33]
 - 4.1e [Machine Thought, Misc](#) [23]
- 4.2 [Computation and Representation](#) [44]
 - 4.2a [Symbols and Symbol Systems](#) [9]
 - 4.2b [Computational Semantics](#) [16]
 - 4.2c [Implicit/Explicit Representation](#) [8]
 - 4.2d [AI without Representation?](#) [6]
 - 4.2e [Miscellaneous](#) [6]
- 4.3 [Philosophy of Connectionism](#) [150]
 - 4.3a [Connectionism and Compositionality](#) (Fodor/Pylyshyn) [43]
 - 4.3b [Representation in Connectionism](#) [20]
 - 4.3c [Connectionism and Eliminativism](#) [16]
 - 4.3d [The Connectionist/Classical Debate](#) [20]
 - 4.3e [Subsymbolic Computation](#) (Smolensky) [8]
 - 4.3f [Philosophy of Connectionism, Misc](#) [36]
 - 4.3g [Foundational Empirical Issues](#) [12]
- 4.4 [Dynamical Systems](#) [22]
- 4.5 [Foundational Questions in AI](#) [51]
 - 4.5a [The Nature of AI](#) [7]
 - 4.5b [Levels of Analysis](#) (Marr, etc) [12]
 - 4.5c [The Frame Problem](#) [14]

- 4.5d [Methodology](#) [19]
- 4.6 [Computationalism in Cognitive Science](#) [33]
- 4.7 [Computation and Physical Systems](#) [18]
- 4.8 [Philosophy of AI, Misc](#) [17]

Part 5: Philosophy of Psychology [561]

- 5.1 [Issues in Cognitive Science](#) [230]
 - 5.1a [Nativism](#) (Chomsky, etc) [37]
 - 5.1b [Modularity](#) (Fodor, etc) [30]
 - 5.1c [Mental Imagery](#) [45]
 - 5.1d [Rationality](#) [31]
 - 5.1e [Embodiment](#) [31]
 - 5.1f [Animal Cognition](#) [56]
- 5.2 [Aspects of Mind](#) [224]
 - 5.2a [Pain and Pleasure](#) [40]
 - 5.2b [Emotions](#) [46]
 - 5.2c [Dreams](#) [21]
 - 5.2d [Memory](#) [32]
 - 5.2e [Color](#) [85]
- 5.3 [Philosophy of Psychology and Cognitive Science](#) [80]
 - 5.3a [Psychological Laws](#) [13]
 - 5.3b [Psychology and Neuroscience](#) [33]
 - 5.3c [Explanation in Cognitive Science](#) [17]
 - 5.3d [Philosophy of Cognitive Science, Misc](#) [17]
- 5.4 [Philosophy of Mind](#) (General) [27]

Part 6: Consciousness in Science [1599]

- 6.1 [Consciousness and Neuroscience](#) [460]
 - 6.1a [Neural Correlates of Consciousness](#) [74]
 - 6.1b [Neural Correlates of Visual Consciousness](#) [28]
 - 6.1c [Blindsight](#) [50]
 - 6.1d [Other Neuropsychological Disorders](#) [71]
 - 6.1e [Cerebral Hemispheres and Consciousness](#) [45]
 - 6.1f [Neural Timing and Consciousness](#) (Libet) [19]
 - 6.1g [Neural Synchrony and Binding](#) [42]
 - 6.1h [Consciousness and Anesthesia](#) [44]

- 6.1i [Foundational Issues](#) [31]
 - 6.1j [Consciousness and Neuroscience, Misc](#) [56]
- 6.2 [Consciousness and Psychology](#) [787]
 - 6.2a [Cognitive Models of Consciousness](#) [60]
 - 6.2b [Unconscious Perception](#) [73]
 - 6.2c [Unconscious Processes](#) [45]
 - 6.2d [Visual Consciousness](#) [37]
 - 6.2e [Consciousness and Attention](#) [35]
 - 6.2f [Consciousness and Memory](#) [84]
 - 6.2g [Consciousness and Learning](#) [33]
 - 6.2h [Consciousness and Metacognition](#) [22]
 - 6.2i [Consciousness and Control](#) [38]
 - 6.2j [Consciousness and Imagery](#) [24]
 - 6.2k [Consciousness and Emotion](#) [31]
 - 6.2l [Consciousness, Sleep, and Dreaming](#) [34]
 - 6.2m [Consciousness in Psychoanalysis](#) [57]
 - 6.2n [Consciousness and Time](#) [35]
 - 6.2o [Self-Consciousness](#) [46]
 - 6.2p [Development of Consciousness](#) [33]
 - 6.2q [The Stream of Consciousness](#) [29]
 - 6.2r [Foundational Issues](#) [19]
 - 6.2s [Consciousness and Psychology, Misc](#) [52]
- 6.3 [Consciousness and Physics](#) [77]
 - 6.3a [Interpretation of Quantum Mechanics](#) [27]
 - 6.3b [Quantum Mechanisms of Consciousness](#) [30]
 - 6.3c [Consciousness and Physics, Misc](#) [20]
- 6.4 [Consciousness and Science, Misc](#) [275]
 - 6.4a [Evolution of Consciousness](#) [31]
 - 6.4b [Consciousness and Language](#) [23]
 - 6.4c [Animal Consciousness](#) [33]
 - 6.4d [Animal Self-Consciousness](#) [22]
 - 6.4e [Altered States of Consciousness](#) [27]
 - 6.4f [Parapsychology and Consciousness](#) [54]
 - 6.4g [Phenomenology](#) [34]
 - 6.4h [Foundations](#) [27]
 - 6.4i [Miscellaneous](#) [24]

Go to:

Part 1: Consciousness and Qualia [1082]

Part of [Contemporary Philosophy of Mind: An Annotated Bibliography](#).

Compiled by [David J. Chalmers](#), Department of Philosophy, University of Arizona, Tucson AZ 85721. E-mail: chalmers@arizona.edu.

- 1.1 [Consciousness](#) [85]
 - 1.1a [General](#) [47]
 - 1.1b [The Concept of Consciousness](#) [38]
- 1.2 [Explaining Consciousness?](#) [115]
 - 1.2a [Subjectivity and Objectivity](#) (Nagel) [46]
 - 1.2b [The Explanatory Gap](#) (Levine) [16]
 - 1.2c [`Hard' and `Easy' Problems](#) (Chalmers) [25]
 - 1.2d [Cognitive Closure](#) (McGinn) [16]
 - 1.2e [Miscellaneous](#) [12]
- 1.3 [Materialism and Dualism](#) [213]
 - 1.3a [The Knowledge Argument](#) (Jackson) [46]
 - 1.3b [Zombies & Modal Arguments](#) [38]
 - 1.3c [Essentialist Arguments](#) (Kripke) [31]
 - 1.3d [Arguments from Disembodiment](#) [17]
 - 1.3e [Consciousness and Physicalism, Misc](#) [14]
 - 1.3f [Consciousness and Dualism](#) [16]
 - 1.3g [Mind-Body Problem, Misc](#) [51]
- 1.4 [Specific Views on Consciousness](#) [227]
 - 1.4a [Higher-Order Thought Approaches](#) (Rosenthal, etc) [46]
 - 1.4b [Dennett on Consciousness](#) [60]
 - 1.4c [Functionalism](#) [8]
 - 1.4d [Eliminativism](#) [16]
 - 1.4e [Epiphenomenalism](#) [30]
 - 1.4f [Interactionism](#) [16]
 - 1.4g [Panpsychism](#) [24]
 - 1.4h [Intrinsic Monism](#) (Russell, Feigl, etc) [27]
- 1.5 [Consciousness and Content](#) [129]
 - 1.5a [Consciousness and Intentionality](#) (Searle, etc) [23]
 - 1.5b [The Content of Experience](#) [48]
 - 1.5c [Representationalism](#) [38]
 - 1.5d [Internalism and Externalism about Experience](#) [12]

- 1.5e [Miscellaneous](#) [8]
- 1.6 [Aspects of Consciousness](#) [182]
 - 1.6a [Self-Consciousness](#) [54]
 - 1.6b [Unity of Consciousness](#) [45]
 - 1.6c [Homogeneity of Consciousness](#) (Sellars, etc) [11]
 - 1.6d [Knowledge of Consciousness](#) [50]
 - 1.6e [The Function of Consciousness](#) [22]
- 1.7 [Qualia](#) [131]
 - 1.7a [Qualia, General](#) [27]
 - 1.7b [Qualia and Materialism](#) [13]
 - 1.7c [Eliminativism about Qualia](#) [11]
 - 1.7d [The Inverted Spectrum](#) [31]
 - 1.7e [Absent Qualia](#) (Block, etc) [15]
 - 1.7f [Introspection and Absent Qualia](#) (Shoemaker) [10]
 - 1.7g [Functionalism and Qualia, General](#) [24]

1.1 Consciousness

1.1a

Consciousness -- General

Armstrong, D. M. & Malcolm, N. 1984. *Consciousness and Causality: A Debate on the Nature of Mind*. Blackwell.

Block, N. , Flanagan, O. , & Guzeldere, G. (eds) 1997. *The Nature of Consciousness: Philosophical Debates*. MIT Press.

An anthology of central philosophical papers on consciousness.

Carruthers, P. 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.

Catalano, J. 2000. *Thinking Matter: Consciousness from Aristotle to Putnam and Sartre*. Routledge.

Chalmers, D. J. 1991. Consciousness and cognition. Manuscript.

Exploring the link between consciousness and judgments about consciousness. Coherence between these
=> consciousness depends on the functional but isn't reducible. Toward a dual-aspect theory based on
pattern and information.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Argues against the reductive explanation of consciousness, and for a kind of naturalistic dualism. Moves toward a "fundamental theory" to bridge the gap, and draws out some consequences.

Churchland, P. M. & Churchland, P. S. 1997. Recent work on consciousness: Philosophical, theoretical, and empirical. *Seminars in Neurology* 17:179-86.

Davies, M. & Humphreys, G. 1993. *Consciousness: Philosophical and Psychological Essays*. Blackwell.

A collection of 5 psychological and 8 philosophical essays on consciousness.

Flanagan, O. J. 1991. Consciousness. In *The Science of the Mind*. MIT Press.

On the mysteries of consciousness. Argues with epiphenomenalism, "conscious inessentialism", and the "new mysterians" (Nagel, McGinn). Toward a naturalistic theory, drawing on ideas of Edelman, Calvin, Dennett.

Flanagan, O. J. 1992. *Consciousness Reconsidered*. MIT Press.

Argues that consciousness can be accounted for in a naturalistic framework. With arguments against eliminativism and epiphenomenalism, evidence from neuroscience and psychology, and discussions of the stream and the self.

Flanagan, O. J. & Guzeldere, G. 1997. Consciousness: A philosophical tour. In (M. Ito, Y. Miyashita, & E. T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.

Foss, J. 2000. *Science and the Riddle of Consciousness: A Solution*. Kluwer.

Graham, G. & Horgan, T. 1998. Sensations and grain processes. In (G. Mulhauser, ed) *Evolving Consciousness*. John Benjamins.

Gregory, R. L. 1988. Consciousness in science and philosophy: conscience and con-science. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Guzeldere, G. 1995. Consciousness: what it is, how to study it, what to learn from its history. *Journal of Consciousness Studies* 2:30-51.

A history of the study of consciousness, especially in psychology.

Guzeldere, G. 1995. Problems of consciousness: A perspective on contemporary issues, current debates. *Journal of Consciousness Studies* 2:112-43.

A summary of recent philosophical debates over consciousness, focusing on the "what/where/who/why/how" questions, the explanatory gap, and the stalemate between "essentialist" and "causal" intuitions.

Hannay, A. 1987. The claims of consciousness: A critical survey. *Inquiry* 30:395-434.

Hannay, A. 1990. *Human Consciousness*. Routledge.

Hurley, S. 1998. *Consciousness in Action*. Harvard University Press.

Jackendoff, R. 1987. *Consciousness and the Computational Mind*. MIT Press.

Separates computational mind from phenomenological mind, and studies the former, a third-person approach. The residue is the "Mind-Mind" problem. Consciousness supervenes on an intermediate level of representation. Elegant.

Kirk, R. 1994. *Raw Feeling: A Philosophical Account of the Essence of Consciousness*. Oxford University Press.

Physicalism can explain consciousness in all its glory. Argues against zombies and inverted-spectrum scenarios, and suggests that the explanatory gap can be bridged by an account of directly-active information-processing.

Levine, J. 1997. Recent work on consciousness. *American Philosophical Quarterly* 34:379-404.

Lycan, W. G. 1987. *Consciousness*. MIT Press.

Lycan, W. G. 1996. *Consciousness and Experience*. MIT Press.

Metzinger, T. (ed) 1995. *Conscious Experience*. Ferdinand Schoningh.

An excellent collection of 20 philosophical papers on consciousness.

Murata, J. 1997. Consciousness and the mind-body problem. In (M. Ito, Y. Miyashita, & E. T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.

Nelkin, N. 1996. *Consciousness and the Origins of Thought*. Cambridge University Press.

O'Shaughnessy, B. 2000. *Consciousness and the World*. Oxford University Press.

Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. MIT Press.

Revonsuo, A. & Kamppinen, M. (eds) 1994. *Consciousness in Philosophy and Cognitive Neuroscience*.

Lawrence Erlbaum.

Sayre, K. M. 1969. *Consciousness: A Philosophic Study of Minds and Machines*. Random House.

Seager, W. E. 1999. *Theories of Consciousness: An Introduction and Assessment*. Routledge.

Searle, J. R. 1989. Consciousness, unconsciousness, and intentionality. *Philosophical Topics* 17:193-209.

Argues that the first-person view has been ignored too much in the philosophy of mind. Even unconscious states are only mental by virtue of their potential consciousness.

Searle, J. R. 1992. *The Rediscovery of the Mind*. MIT Press.

On the centrality of consciousness to the mind. Consciousness is irreducible but biological. On the history of the field, the structure of consciousness, its role in constituting intentionality, and problems with computation.

Searle, J. R. 1993. The problem of consciousness. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

On the notion of consciousness, its relation to the brain, and some features that need to be explained: its subjectivity, unity, intentionality, center and periphery, Gestalt structure, aspect of familiarity, and so on.

Sheets-Johnstone, M. 1998. Consciousness: A natural history. *Journal of Consciousness Studies* 5:260-94.

Siewert, C. 1998. *The Significance of Consciousness*. Princeton University Press.

Smith, D. W. 1992. Consciousness in action. *Synthese* 90:119-43.

Sprigge, T. L. S. 1982. The importance of subjectivity: An inaugural lecture. *Inquiry* 25:143-63.

Value is only found within streams of consciousness. Three ways of studying it: phenomenology, anthropology, and by relation to the physical. With an analysis of the "self-transcending" nature of conscious intentionality.

Strawson, G. 1994. *Mental Reality*. MIT Press.

Sturgeon, S. 2000. *Matters of Mind: Consciousness, Reason and Nature*. Routledge.

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press.

Tye, M. 2000. *Consciousness, Color, and Content*. MIT Press.

Velmans, M. 1996. *The Science of Consciousness: Psychological, Neuropsychological, and Clinical Reviews*. Routledge.

Verges, F. G. 1974. Jackson on incorrigibility. *Australasian Journal of Philosophy* 52:243-50.

Villaneuva, E. (ed) 1991. *Consciousness: Philosophical Issues*. Ridgeview.

A collection of philosophical articles on consciousness.

Young, A. W. & Block, N. 1997. Consciousness. In (V. Bruce, ed) *Unsolved Mysteries of the Mind: Tutorial Essays in Cognition*. Taylor and Francis.

1.1b

The Concept of Consciousness

Armstrong, D. M. 1979. Three types of consciousness. In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.

Armstrong, D. M. 1981. What is consciousness? In *The Nature of Mind*. Cornell University Press.

On minimal consciousness, perceptual consciousness, and introspective consciousness. Introspective consciousness seems so special because it gives inner awareness of self, and memory of other mental events.

Baruss, I. 1986. Meta-analysis of definitions of consciousness. *Imagination, Cognition, and Personality* 6:321-29.

Bisiach, E. 1988. The (haunted) brain and consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Distinguishes C1 (phenomenal experience) from C2 (access of parts of a system to other parts). C2 is can be scientifically studied, and has a graspable, if fragmented, causal role. C1 is mysterious and perhaps beyond science.

Burt, C. 1962. The concept of consciousness. *British Journal of Psychology* 53:229-42.

Cam, P. 1985. Phenomenology and speech dispositions. *Philosophical Studies* 47:357-68.

Reportability is not phenomenology, as blindsight has reportability but no phenomenology.

Chalmers, D. J. 1997. Availability: The cognitive basis of experience? In (N. Block, O. Flanagan, and G. Guzeldere, eds) *The Nature of Consciousness*. MIT Press.

Argues that the cognitive correlate of consciousness is direct availability for global control.

Church, J. 1998. Two sorts of consciousness? *Communication and Cognition* 31:51-71.

Finkelstein, D. H. 1999. On the distinction between conscious and unconscious states of mind. *American Philosophical Quarterly* 36:79-100.

Gennaro, R. J. 1995. Does mentality entail consciousness? *Philosophia* 24:331-58.

Girle, R. A. 1996. Shades of consciousness. *Minds and Machines* 6:143-57.

Kirk, R. 1992. Consciousness and concepts. *Aristotelian Society Supplement* 66:23-40.

Analyzes consciousness in terms of the "presence" of information to the main decision-making processes of a system. No great conceptual capacities required, no higher-order thoughts. With application to blindsight.

Lormand, E. 1995. What qualitative consciousness is like. Manuscript.

Lormand, E. 1996. Nonphenomenal consciousness. *Nous* 30:242-61.

Manson, N. 2000. State consciousness and creature consciousness: A real distinction. *Philosophical Psychology* 13:405-410.

Matthews, G. 1977. Consciousness and life. *Philosophy* 52:13-26.

McBride, R. 1999. Consciousness and the state/transitive/creature distinction. *Philosophical Psychology* 12:181-196.

Moody, T. C. 1986. Distinguishing consciousness. *Philosophy and Phenomenological Research* 47:289-95.

Separates consciousness from the mental -- functionalist accounts work for the latter but not the former. With remarks on Zen "pure consciousness".

Natsoulas, T. 1978. Consciousness. *American Psychologist* 33:906-14.

On the role of consciousness in psychology, and distinguishing various notions of consciousness: mutual knowledge, internal knowledge, awareness, direct awareness, personal unity, wakefulness, and double consciousness.

Natsoulas, T. 1983. A selective review of conceptions of consciousness with special reference to behavioristic contributions. *Cognition and Brain Theory* 6:417-47.

Ideas about consciousness from Locke, Brentano, Hebb, Dennett, Skinner, Sellars, Aristotle, Gibson.
Theories: inner eye vs. verbal vs. outer eye.

Natsoulas, T. 1983. Concepts of consciousness. *Journal of Mind and Behavior* 4:195-232.

Natsoulas, T. 1991. The concept of consciousness(1): The interpersonal meaning. *Journal for the Theory of Social Behavior* 21:63-89.

Natsoulas, T. 1991. The concept of consciousness(2): The personal meaning. *Journal for the Theory of Social Behavior* 21:339-67.

Natsoulas, T. 1992. The concept of consciousness(3): The awareness meaning. *Journal for the Theory of Social Behavior* 2:199-25.

Natsoulas, T. 1993. Consciousness(4): Varieties of intrinsic theory. *Journal of Mind and Behavior* 14:107-32.

Natsoulas, T. 1994. The concept of consciousness(4): The reflective meaning. *Journal for the Theory of Social Behavior* 24:373-400.

Natsoulas, T. 1994. The concept of consciousness(5): The unitive meaning. *Journal for the Theory of Social Behavior* 24:401-24.

Natsoulas, T. 1995. Consciousness(3) and Gibson's concept of awareness. *Journal of Mind and Behavior* 3:305-28.

Natsoulas, T. 1996-1998. The case for intrinsic theory (parts 1-3). *Journal of Mind and Behavior* 17:267-85, 17:369-89, 19:1.

Natsoulas, T. 1997. Consciousness and self-awareness: Consciousness(1,2,3,4,5,6). *Journal of Mind and Behavior* 18:53-94.

Nelkin, N. 1987. What is it like to be a person? *Mind and Language* 21:220-41.

Critiques three senses of consc: awareness, verbalization and phenomenology. Argues that none are sufficient for person-consciousness. Quite good.

Nelkin, N. 1993. What is consciousness? *Philosophy of Science* 60:419-34.

On three senses of consciousness: phenomenality, intentionality, and introspectibility. Argues from empirical evidence (especially blindsight cases) that these three are all dissociable.

O'Shaughnessy, B. 1991. The anatomy of consciousness. In (E. Villanueva, ed) *Consciousness*. Ridgeview.

Place, U. T. 1992. Two concepts of consciousness: The biological/private and the linguistic/social. *Acta Analytica* 7:53-72.

Rosenthal, D. M. 1990. The independence of consciousness and sensory quality. In (E. Villanueva, ed) *Consciousness*. Ridgeview.

Argues that consciousness and sensory quality are independent properties: there can be unconscious sensations. Consciousness is a relational property.

Rosenthal, D. M. 1994. State consciousness and transitive consciousness. *Consciousness and Cognition* 2:355-63.

Shanon, B. 1990. Consciousness. *Journal of Mind and Behavior* 11:137-51.

On three kinds of consciousness -- sensed being, mental awareness, and reflection -- and their relationships.

Tye, M. 1996. The burning house. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Uses various puzzles cases to distinguish higher-order consciousness, discriminatory consciousness, responsive consciousness, and phenomenal consciousness.

1.2

Explaining Consciousness?

1.2a

Subjectivity and Objectivity (Nagel) [see also 1.3a]

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 4:435-50. Reprinted in *Mortal Questions* (Cambridge University Press, 1979).

Physicalist explanations leave out consciousness, i.e. what it is like to be an organism. Objective accounts omit points of view (could there be an objective phenomenology?). Physicalism may be true, but we can't

see how.

Nagel, T. 1979. Subjective and objective. In *Mortal Questions*. Cambridge University Press.

Subjective and objective views clash e.g. on meaning of life, free will, personal identity, mind-body problem, ethics. How to reconcile: reduction, elimination, annexation? Maybe just let multiple viewpoints coexist.

Nagel, T. 1986. *The View From Nowhere*. Oxford University Press.

Seeing philosophy as a clash between the subjective and objective views of various phenomena (mental states, self, knowledge, freedom, value, ethics). Eliminating the subjective is impossible.

Akins, K. 1993. What is it like to be boring and myopic? In (B. Dahlbom, ed) *Dennett and his Critics*. Blackwell.

Gives a detailed account of perceptual processing in bats, and suggests that we can know what bat-experience is like: it's like nerd experience. But then is there an unexplained residue?

Akins, K. 1993. A bat without qualities? In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

On what science tells us about the experience of bats, birds, and others. Why a movie of bat-experience isn't good enough -- because of the inseparability of intentionality and experience. Science can do OK.

Baker, L. R. 1998. The first-person perspective: A test for naturalism. *American Philosophical Quarterly* 35:327-348.

Biro, J. I. 1991. Consciousness and subjectivity. In (E. Villanueva, ed) *Consciousness*. Ridgeview.

No real problems are posed by subjectivity and points of view, no matter how they are construed (fixed, portable, tokens, types). It's either a confusion or a triviality about the logic of indexicality.

Biro, J. I. 1993. Consciousness and objectivity. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Carruthers, P. 2000. Sympathy and subjectivity. *Australasian Journal of Philosophy* 77:465-482.

Davis, L. 1982. What is it like to be an agent? *Erkenntnis* 18:195-213.

On what is required for consciousness of agency (rather than qualia): belief, intention, and most importantly desire, enabling a capacity to care. A robot could have all this, and it would be like something to be it.

Flanagan, O. J. 1985. Consciousness, naturalism and Nagel. *Journal of Mind and Behavior* 6:373-90.

Naturalism can do autophenomenology just fine.

Foss, J. E. 1989. On the logic of what it is like to be a conscious subject. *Australasian Journal of Philosophy* 67:305-320.

A Super Neuroscientist will know how we describe and think about experience, so will know as much as a Super Sympathist. One doesn't have to imagine to know what it's like. With remarks on bat experience.

Foss, J. E. 1993. Subjectivity, objectivity, and Nagel on consciousness. *Dialogue* 32:725-36.

Nagel conflates metaphysical and epistemological versions of the subjective/ objective distinction. Consciousness is metaphysically subjective, and science is epistemically objective, so there is incompatibility.

Francescotti, R. M. 1993. Subjective experience and points of view. *Journal of Philosophical Research* 18:25-36.

Being graspable from only one point of view does not define the class of facts about conscious experience. Various ways of cashing this out fail.

Haksar, V. 1981. Nagel on subjective and objective. *Inquiry* 24:105-21.

The objective and subjective don't conflict, but complement each other.

Hanna, P. 1990. Must thinking bats be conscious? *Philosophical Investigations* 13:350-55.

Hiley, D. R. 1978. Materialism and the inner life. *Southern Journal of Philosophy* 16:61-70.

Nagel conflates questions about sensory qualities with those about a unique point of view. The truth of physicalism is irrelevant to uniqueness.

Hill, C. S. 1977. Of bats, brains, and minds. *Philosophy and Phenomenological Research* 38:100-106.

Kekes, J. 1977. Physicalism and subjectivity. *Philosophy and Phenomenological Research* 37:533-6.

The subjective/objective distinction is ill-drawn. Objective descriptions aren't species-independent, but in terms of the space-time causal network. Science can explain the experience this way, but not provide the experience.

Lewis, D. 1983. Postscript to "Mad pain and Martian pain". In *Philosophical Papers*, Vol. 1. Cambridge University Press.

Knowing what it's like consists in an ability, not possession of information.

Lycan, W. G. 1987. "Subjectivity". In *Consciousness*. MIT Press.

Various anti-Nagel points.

Lycan, W. G. 1990. What is the "subjectivity" of the mental? *Philosophical Perspectives*.

The subjectivity of the mental is no more special than usual propositional subjectivity. It can be handled by a self-scanner model of introspection.

Malcolm, N. 1988. Subjectivity. *Philosophy* 63:147-60.

A critique of Nagel's idea of a "point of view" that is occupied by a "subject". There aren't any peculiar facts about given viewpoints.

Maloney, J. C. 1986. About being a bat. *Australasian Journal of Philosophy* 64:26-49.

Mandik, P. 2001. Mental representation and the subjectivity of consciousness. *Philosophical Psychology* 14:179-202.

Mellor, D. H. 1993. Nothing like experience. *Proceedings of the Aristotelian Society* 63:1-16.

There are no fact about what an experience is like. Knowing what it's like is an ability to imagine, recognize, and recall; this explains ineffability, etc. With remarks on the experience of imagining an experience.

McClamrock, R. 1992. Irreducibility and subjectivity. *Philosophical Studies* 67:177-92.

Phenomenological properties cannot be picked out in physical or computational terms; argues against Lycan's criticism of Nagel. But all this is compatible with materialism. With comments on the phenomenological tradition.

McCulloch, G. 1988. What it is like. *Philosophical Quarterly* 38:1-19.

Criticizes absent/inverted qualia arguments for a special "what it is like", but argues that the possibility of "what it is like" differences relative to semantic states shows that something's not conveyed by functional accounts.

McMullen, C. 1985. 'Knowing what it's like' and the essential indexical. *Philosophical Studies* 48:211-33.

The Nagel/Jackson argument is analogous to the Perry indexical argument, and can be treated the same way.

Mounce, H. O. 1992. On Nagel and consciousness. *Philosophical Investigations* 15:178-84.

Muscari, P. 1985. The subjective character of experience. *Journal of Mind and Behavior* 6:577-97.

Muscari, P. 1987. The status of humans in Nagel's phenomenology. *Philosophical Forum* 19:23-33.

Nagel's dilemma: separating feeling from process. Moral consequences?

Nelkin, N. 1987. What is it like to be a person? *Mind and Language* 2:220-41.

Nagel-consciousness exists, but isn't so important. It's essential for sensations, but not for thoughts. Beings without it could still be persons.

Nemirow, L. 1980. Review of Nagel's *Mortal Questions*. *Philosophical Review* 89:473-7.

Understanding does not consist only in facts; we can understand via sympathy.

Nemirow, L. 1990. Physicalism and the cognitive role of acquaintance. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Knowing what it's like is really knowing how to imagine. We should reduce Nagel's question to a question about possession of a certain ability.

Pugmire, D. 1989. Bat or batman. *Philosophy* 64:207-17.

Subjectivity is not something we have knowledge of, as we lack comparisons.

Rorty, R. 1993. Holism, intrinsicity, and the ambition of transcendence. In (B. Dahlbom, ed) *Dennett and His Critics*. Blackwell.

On the Nagel/Dennett debate: Nagel holds out for unexplained intrinsic properties once the relational is all accounted for; Dennett can renounce the transcendental ambition. Remarks on realism, holism, and metaphilosophy.

Rudd, A. J. 1999. What it's like and what's really wrong with physicalism: A Wittgensteinian perspective. *Journal of Consciousness Studies* 5:454-63.

Russow, L. 1982. It's not like that to be a bat. *Behaviorism* 10:55-63.

Divides Nagel's problem: qualitative differences, special access, mineness.

Simoni-Wastila, H. 2000. Particularity and consciousness: Wittgenstein and Nagel on privacy, beetles and

bats. *Philosophy Today* 44:415-425.

Taliaferro, C. 1988. Nagel's vista or taking subjectivity seriously. *Southern Journal of Philosophy* 26:393-401.

Nagel's 'View from Nowhere' doesn't take subjectivity seriously enough.

Teller, P. 1992. Subjectivity and knowing what it's like. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Rebutting various intuitions for the non-physical nature of experience. The Nagel/Jackson argument commits an intensional fallacy; experiences are physical states known from a different perspective.

Tilghman, B. R. 1991. What is it like to be an aardvark? *Philosophy* 66:325-38.

A Wittgensteinian critique of Nagel. Nagel's question is confused: "what it's like" is a matter of behavior, sociality, etc, not inner experience.

van Gulick, R. 1985. Physicalism and the subjectivity of the mental. *Philosophical Topics* 13:51-70.

Reducing doesn't imply understanding. Two different kinds of reduction.

Wider, K. 1989. Overtones of solipsism in Nagel's 'What is it like to be a bat?' and 'The view from nowhere'. *Philosophy and Phenomenological Research* 49:481-99.

Nagel is an epistemological solipsist, whether he likes it or not.

Wright, E. 1996. What it isn't like. *American Philosophical Quarterly* 33:23-42.

1.2b The Explanatory Gap (Levine)

Beckermann, A. 2000. The perennial problem of the reductive explainability of phenomenal consciousness: C. D. Broad on the explanatory gap. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Bieri, P. 1995. Why is consciousness puzzling? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Reflections on the explanatory gap between physical processes and conscious experience. With remarks on different sorts of consciousness, and on why we need an intelligible necessary connection.

Block, N. & Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical*

Review.

Chalmers, D. J. & Jackson, F. 2001. Conceptual analysis and reductive explanation. *Philosophical Review*.

Ellis, R. D. & Newton, N. 1998. Three paradoxes of phenomenal consciousness: Bridging the explanatory gap. *Journal of Consciousness Studies* 5:419-42. Hardin, C. L. 1987. Qualia and materialism: Closing the explanatory gap. *Philosophy and Phenomenological Research* 48:281-98.

On the physiological bases of phenomenal states, particularly color. Inverted spectrum isn't really coherent, as coolness/warmth would have to be inverted too. So the contingency of qualia is diminished.

Hardin, C. L. 1992. Physiology, phenomenology, and Spinoza's true colors. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Argues in detail that psychophysics can provide a structural map to close the explanatory gap. If there is an explanatory residue, perhaps panpsychism can help.

Harnad, S. 1994. Why and how we are not zombies. *Journal of Consciousness Studies* 1:164-67.

Kim, J. 1998. Reduction, reductive explanation, and "the explanatory gap". Manuscript.

Argues for a distinction between reduction and reductive explanation, and argues that reductive explanations generally involves conceptual connections via "functionalization". With comments on Block and Stalnaker.

Levin, J. 1991. Analytic functionalism and the reduction of phenomenal states. *Philosophical Studies* 61:211-38.

Contra Kripkean arguments, a good enough functional theory may help close the conceivability/explanatory gap between the physical and qualia. Contra Nagel/Jackson, such a theory could provide us with recognitional abilities.

Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.

How do we explain the apparent contingency of the qualia-matter reduction? Even if it's not metaphysically contingent, it's conceptually contingent, so there's a gap in any physical explanation of qualia. Excellent.

Levine, J. 1993. On leaving out what it's like. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Physical accounts leave out qualia epistemologically but not metaphysically. So physicalism holds, but there is an explanatory gap. Discusses Kripke's and Jackson's arguments in detail; also explanation and

content.

Papineau, D. 1998. Mind the gap. *Philosophical Perspectives* 12:373-89.

Price, M. C. 1996. Should we expect to feel as if we understand consciousness? *Journal of Consciousness Studies* 3:303-12.

Argues that the explanatory gap between brain and consciousness is just the same as that found with causal relations everywhere; it's just that we usually overlook it in the latter case.

Sturgeon, S. 1994. The epistemic basis of subjectivity. *Journal of Philosophy* 91:221-35.

Qualia can't be explained in more basic terms, as that sort of explanation works by accounting for a property's canonical evidence, but the canonical evidence for qualia are qualia themselves. But they still may be physical.

Tye, M. 1999. Phenomenal consciousness: The explanatory gap as a cognitive illusion. *Mind* 108:705-25.

1.2c `Hard' and `Easy' Problems (Chalmers) [see also 3.1]

Chalmers, D. J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:200-19. Also in (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press. Reprinted in Shear 1997.

Divides the problems of consciousness into easy and hard problems; the hard problem eludes reductive explanation as it isn't about explaining functions. Argues instead for a nonreductive theory with psychophysical laws.

Chalmers, D. J. 1995. The puzzle of conscious experience. *Scientific American* 273(6):80-86.

Like the JCS article, but shorter, more accessible, and with pretty pictures.

Chalmers, D. J. 1996. Can consciousness be reductively explained? In *The Conscious Mind*. Oxford University Press.

There is no a priori entailment from the physical to phenomenal facts (arguments from conceivability, epistemology, analysis), so reductive explanation fails. With a critique of existing empirical proposals.

Chalmers, D. J. 1997. Moving forward on the problem of consciousness. *Journal of Consciousness Studies* 4:3-46. Reprinted in Shear 1997.

A reply to the 25 "hard problem" articles in JCS.

Chalmers, D. J. 1998. The problems of consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Churchland, P. S. 1996. The hornswoggle problem. *Journal of Consciousness Studies* 3:402-8. Reprinted in Shear 1997,

Argues that the "hard problem" in effect invokes an argument from ignorance, and that there's no deep difference between consciousness and other domains.

Clark, T. 1995. Function and phenomenology: Closing the explanatory gap. *Journal of Consciousness Studies* 2:241-54. Reprinted in Shear 1997.

Argues contra Chalmers that experience is identical to certain functions, rather than emerging from them.

Crick, F. and Koch, C. 1995. Why neuroscience may be able to explain consciousness. *Scientific American* 273(6):84-85. Reprinted in Shear 1997.

Divides the hard problem into three parts, and argues that neuroscience can make progress on at least one part (incommunicability); and maybe "meaning" holds the key to the rest.

Dennett, D. C. 1996. Facing backwards on the problem of consciousness. *Journal of Consciousness Studies* 3:4-6. Reprinted in Shear 1997.

Argues contra Chalmers 1995 that functions are all we need to explain.

Eilan, N. 2000. Primitive consciousness and the 'hard problem'. *Journal of Consciousness Studies* 7:28-39.

Gray, J. 1998. Creeping up on the hard question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Hodgson, D. 1996. The easy problems ain't so easy. *Journal of Consciousness Studies* 3:69-75. Reprinted in Shear 1997.

Argues that consciousness plays a vital role in performing mental functions, so the easy problems won't be solved until the hard problem is solved.

Horst, S. 1999. Evolutionary explanation and the hard problem of consciousness. *Journal of Consciousness Studies* 6:39-48.

Ismael, J. 1999. Science and the phenomenal. *Philosophy of Science* 66:351-69.

Lewis, H. 1998. Consciousness: Inexplicable - and useless too? *Journal of Consciousness Studies* 5:59-66.

Libet, B. 1996. Solutions to the hard problem of consciousness. *Journal of Consciousness Studies* 3:33-35. Reprinted in Shear 1997.

Endorses the idea of consciousness as fundamental, but criticizes Chalmers' psychophysical laws. Advocates a theory with a "conscious mental field".

Lowe, E. J. 1995. There are no easy problems of consciousness. *Journal of Consciousness Studies* 2:266-71. Reprinted in Shear 1997.

Argues that the "easy problems" -- reportability, attention, etc -- all involve concepts and therefore experience itself, for Kantian reasons, and therefore are not mechanistically explainable.

Mills, E. O. 1996. Giving up on the hard problem of consciousness. *Journal of Consciousness Studies* 3:26-32. Reprinted in Shear 1997.

Argues that the truly hard problem is that of giving a constitutive account of consciousness, and Chalmers doesn't solve that (laws aren't good enough); in fact it's unsolvable.

Mills, F. B. 1998. The easy and hard problems of consciousness: A Cartesian perspective. *Journal of Mind and Behavior* 19:119-40.

O'Hara, K. & Scutt, T. 1996. There is no hard problem of consciousness. *Journal of Consciousness Studies* 3:290-302. Reprinted in Shear 1997.

Argue that we should work on the easy problems for now, as nobody has any good ideas about the hard problem; maybe it will gradually fade away.

Robinson, W. S. 1996. The hardness of the hard problem. *Journal of Consciousness Studies* 3:14-25. Reprinted in Shear 1997.

Shear, J. 1996. The hard problem: Closing the empirical gap. *Journal of Consciousness Studies* 3:54-68. Reprinted in Shear 1997.

On the epistemology of the hard problem. Argues that a scientific study of phenomenology is possible, drawing on work in developmental psychology and Eastern thought. "Pure consciousness" may be relevant to a resolution.

Shear, J. (ed) 1997. *Explaining Consciousness: The Hard Problem*. MIT Press.

A collection of essays consisting of Chalmers' keynote paper, 26 replies from many perspectives, and Chalmers' response to the replies.

Varela, F. 1995. Neurophenomenology: A methodological remedy for the hard problem. *Journal of*

Consciousness Studies 3:330-49. Reprinted in Shear 1997.

Advocates a careful phenomenological study of consciousness in its own right, systematically linked with a neurophysiological investigation.

Velmans, M. 1995. The relation of consciousness to the material world. *Journal of Consciousness Studies* 2:255-65. Reprinted in Shear 1997.

Agrees with Chalmers on nonreductionism, but disagrees on "awareness", organizational invariance, and thermostats. Advocates a kind of dual-aspect theory, where the physical world is present within consciousness,

1.2d Cognitive Closure (McGinn)

Davies, W. M. 1999. Sir William Mitchell and the "new mysterianism". *Australasian Journal of Philosophy* 77:253-73.

Garvey, J. 1997. What does McGinn think we cannot know? *Analysis* 57:196-201.

Hanson, P. P. 1993. McGinn's cognitive closure. *Dialogue* 32:579-85.

Kirk, R. 1991. Why shouldn't we be able to solve the mind-body problem? *Analysis* 51:17-23.

McGinn asks too much of a solution to the M-B problem. We might understand consciousness without understanding specific experiences; we could get at it by studying brain and consciousness not separately but simultaneously.

Krellenstein, M. F. 1995. Unsolvable problems, visual imagery, and explanatory satisfaction. *Journal of Mind and Behavior* 16:235-54.

Kukla, A. 1995. Mystery, mind, and materialism. *Philosophical Psychology* 8:255-64.

McDonough, R. 1992. The last stand of mechanism. *Journal of Speculative Philosophy* 6:206-25.

McGinn, C. 1989. Can we solve the mind-body problem? *Mind* 98:349-66. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Argues that the mind-body problem might be solvable in principle, but beyond human capacities. Neither perception of the brain nor introspection of consciousness can uncover the property by which consciousness arises.

McGinn, C. 1991. *The Problem of Consciousness: Essays Toward a Resolution*. Blackwell.

A collection of articles on the problem of consciousness, advocating a view on which the phenomenon is natural but permanently mysterious to us.

McGinn, C. 1991. Consciousness and the natural order. In *The Problem of Consciousness*. Blackwell.

Argues that a naturalistic account of the intentionality of conscious states requires an account of their embodiment; and that embodiment may depend on the hidden structure of conscious states, not accessible to introspection.

McGinn, C. 1991. The hidden structure of consciousness. In *The Problem of Consciousness*. Blackwell.

Suggests that consciousness may have a hidden structure, analogous to the deep structure of language, that relates its surface properties to physical properties. We may not be able to understand this hidden structure, however.

McGinn, C. 1993. *Problems in Philosophy*. Blackwell.

McGinn, C. 1995. Consciousness and space. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

McGinn, C. 1999. *The Mysterious Flame: Conscious Minds in a Material World*. Basic Books.

Sacks, M. 1994. Cognitive closure and the limits of understanding. *Ratio* 7:26-42

Whitely, C. H. 1990. McGinn on the mind-body problem. *Mind* 99:289.

1.2e Miscellaneous

Churchland, P. M. 1996. The rediscovery of light. *Journal of Philosophy* 93:211-28.

Parodies arguments by Searle, Jackson, and Chalmers for the irreducibility of consciousness with analogous arguments for the irreducibility of "luminescence". The consciousness arguments are no better.

Churchland, P. S. 1998. What should we expect from a theory of consciousness? In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Hardcastle, V. G. 1993. The naturalists versus the skeptics: The debate over a scientific understanding of consciousness. *Journal of Mind and Behavior* 14:27-50.

Argues that consciousness can be handled within a scientific framework. We can translate first-person accounts into third-person accounts. Replies to skeptical objections using analogies from elsewhere in science.

Hardcastle, V. G. 1996. The why of consciousness: A non-issue for materialists. *Journal of Consciousness Studies* 3:7-13.

A "committed materialist" will not see any explanatory gap, or any "brute fact". The entrenched differences lie in one's choice of initial framework.

Hesslow, G. 1996. Will neuroscience explain consciousness? *Journal of Theoretical Biology* 171:29-39.

Kirk, R. 1995. How is consciousness possible? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Kurthen, M. 1995. On the prospects of a naturalistic theory of phenomenal consciousness. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Lockwood, M. 1998. The enigma of sentience. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Nida-Rumelin, M. 1997. Is the naturalization of qualitative experience possible or sensible? In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Taylor, J. G, 1998. Cortical activity and the explanatory gap. *Consciousness and Cognition* 7:109-48.

van Gulick, R. 1993. Understanding the phenomenal mind: Are we all just armadillos? In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Qualia pose no insurmountable problems for materialism: knowledge argument can be answered, explanatory gap can be closed, and absent qualia arguments beg the question. With speculations on their functional role.

van Gulick, R. 1995. What would count as explaining consciousness? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Distinguishes six explananda, four explanatory restrictions, and four sorts of relations between them, making 96 possible problems. With a discussion of whether and how the central problems might be answered.

1.3

Materialism and Dualism

1.3a

The Knowledge Argument (Jackson) [see also 1.2a]

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Knowing a completed neuroscience does not imply knowing about qualia. Mary, the colorblind neuroscientist, gains color vision and learns about red. So physicalism is false, as there are facts over and above the physical facts.

Jackson, F. 1986. What Mary didn't know. *Journal of Philosophy* 83:291-5.

Reply to Churchland 1985: Mary *learns*, Churchland misstates the argument.

Alter, T. 1995. Mary's new perspective. *Australasian Journal of Philosophy* 73:585-84.

Contra Pereboom 1994: The way a color sensation appears is a fact about it.

Alter, T. 1998. A limited defense of the knowledge argument. *Philosophical Studies* 90:35-56.

Bachrach, J. E. 1990. Qualia and theory reduction: A criticism of Paul Churchland. *Iyyun* 281-94.

Argues that Churchland's neuroscientific descriptions must leave at least some qualia behind: they might account for what we know (e.g. brain states) in qualia-knowledge, but can't handle distinctions in how we know.

Bigelow, J. & Pargetter, R. 1990. Acquaintance with qualia. *Theoria*.

Mary gains knowledge of old facts, in a new way: she gains a new mode of acquaintance with those facts. Analogies with indexical knowledge: her new knowledge eliminates no possible worlds.

Churchland, P. M. 1985. Reduction, qualia and the direct introspection of brain states. *Journal of Philosophy* 82:8-28. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Qualia can undergo a normal reduction to the neurophysiological. Jackson commits an intensional fallacy; in any case, perhaps Mary can understand red. When we apprehend qualia, we are directly introspecting our brain state.

Churchland, P. M. 1989. Knowing qualia: A reply to Jackson. In *A Neurocomputational Perspective*. MIT Press.

Rejoinder to Jackson 1986. The key lies in knowing-how vs. knowing-that.

Conee, E. 1985. Physicalism and phenomenal properties. *Philosophical Quarterly* 35:296-302.

Contra Lewis, Nemirow, and Horgan on the knowledge argument. But qualia may still be physical (though outside vocab of science) due to their causal role.

Conee, E. 1994. Phenomenal knowledge. *Australasian Journal of Philosophy*.

Mary knew all the facts about qualia beforehand, she just wasn't acquainted with them. Knowledge by acquaintance isn't factual knowledge.

Cummins, R. 1984. The mind of the matter: Comments on Paul Churchland. *Philosophy of Science Association* 1984, 2:791-8.

Speculation on how consciousness might be left out by a physical account.

Dennett, D. C. 1991. "Epiphenomenal" qualia? In *Consciousness Explained*, pp. 398-406. Little-Brown.

Argues that most people don't really imagine Mary's situation. In fact, Mary would be able to identify blue objects from the way they make her react.

Furash, G. 1989. Frank Jackson's knowledge argument against materialism. *Dialogue* 32:1-6.

Defends Jackson's argument against criticisms by Nemirow, Smith & Jones, Warner, Horgan, & Conee. The argument forces physicalism into a quandary: either deny qualia, or make the confused claim that qualia are physical.

Gertler, B. 1999. A defense of the knowledge argument. *Philosophical Studies* 93:317-336.

Graham, G. & Horgan, T. 2000. Mary Mary, quite contrary. *Philosophical Studies* 99:59-87.

Harman, G. 1993. Can science understand the mind? In (G. Harman, ed) *Conceptions of the Human Mind: Essays on Honor of George A. Miller*. Lawrence Erlbaum.

On Dilthey's "Verstehen", or "understanding from within". Mostly about meaning, but with application to the knowledge Mary gains.

Hershfield, J. 1998. Lycan on the subjectivity of the mental. *Philosophical Psychology* 11:229-38.

Horgan, T. 1984. Jackson on physical information and qualia. *Philosophical Quarterly* 34:147-83.

Mary didn't know all the physical facts: she knew all the explicitly physical information, but not all the ontologically physical information.

Jacquette, D. 1995. The blue banana trick: Dennett on Jackson's color scientist. *Theoria* 61:217-30.

Kelly, J. S. 1989. On neutralizing introspection: The data of sensuous awareness. *Southern Journal of Philosophy* 27:29-53.

Lahav, R. 1994. A new challenge for the physicalist: Phenomenal indistinguishability. *Philosophia* 24:77-103.

A new version of the knowledge argument: given all the physical facts, one can't know when two experiences are indistinguishable. This avoids various objections to the standard version.

Levin, J. 1986. Could love be like a heatwave?: Physicalism and the subjective character of experience. *Philosophical Studies* 49:245-61. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Contra Nagel/Jackson: Understand qualia through relational properties, and separate the mental concept from the recognitional capacity.

Lewis, D. 1990. What experience teaches. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Against the hypothesis that phenomenology carries information. If it does, then qualia are epiphenomenal. Better to analyze the "new information" as acquiring an ability instead. In-depth and entertaining.

Loar, B. 1990. Phenomenal states. *Philosophical Perspectives* 4:81-108.

Phenomenal and functional concepts are distinct, but the relevant properties may be identical. We directly refer to phenomenal properties by recognition. Remarks on other minds, transparency, incorrigibility & more. A meaty paper.

Lycan, W. G. 1995. A limited defense of phenomenal information. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Gives nine arguments against the Lewis/Nemirow ability analysis, and proposes instead that the same fact is learned a new way, like water and H₂O facts. This sort of phenomenal information is no danger to materialism.

Lycan, W. G. 1998. Phenomenal information again: It is both real and intrinsically perspectival. *Philosophical Psychology* 11:239-42.

McConnell, J. 1995. In defense of the knowledge argument. *Philosophical Topics* 22:157-187.

Defends against objections from Dennett, Churchland, etc. Horgan's objection (same fact different ways) has a certain force, but the argument can be reformulated to avoid them and imply property dualism. With remarks on Loar.

Nemirow, L. 1995. Understanding rules. *Journal of Philosophy* 92:28-43.

Newton, N. 1986. Churchland on direct introspection of brain states. *Analysis* 46:97-102.

Contra Churchland 1985: we couldn't introspect sensations as brain states, although we could interpret them as such.

Nida-Rumelin, M. 1995. What Mary couldn't know: Belief about phenomenal states. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Reconstructing the Mary case as a Marianna case, and introducing a distinction between phenomenal and nonphenomenal beliefs, to support the knowledge argument. A very nice and sophisticated paper.

Nida-Rumelin, M. 1998. On belief about experiences: An epistemological distinction applied to the knowledge argument against physicalism. *Philosophy and Phenomenological Research* 58:51-73.

Papineau, D. 1993. Physicalism, consciousness, and the antipathetic fallacy. *Australasian Journal of Philosophy* 71:169-83.

Mary goes from a third-person concept of experience to a first-person concept, but they co-refer; we can refer to an experience without having the experience. Physical and phenomenal properties are brutally identical.

Papineau, D. 1995. The antipathetic fallacy and the boundaries of consciousness. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Pereboom, D. 1994. Bats, brain scientists, and the limits of introspection. *Philosophy and Phenomenological Research* 54:315-29.

Mary learns an old fact under a new mode of presentation, and doesn't even learn a new fact about a mode of presentation. Her access to internal states is always mediated by representation, so we can always ascend to a new mode.

Raymont, P. 1995. Tye's criticism of the knowledge argument. *Dialogue* 34:713-26.

Raymont, P. 1999. The know-how response to Jackson's knowledge argument. *Journal of Philosophical Research* 24:113-26.

Robinson, D. 1993. Epiphenomenalism, laws, and properties. *Philosophical Studies* 69:1-34.

A thorough discussion of Lewis 1990. Phenomenal information implies epiphenomenalism, even at the intra-psychic level. Remarks on ineffability, and on whether properties should be individuated by nomic role or by essence.

Robinson, H. 1993. Dennett on the knowledge argument. *Analysis* 53:174-7.

Contra Dennett, Mary can't tell an object's color unless she already knows about experience. The knowledge argument bears on thought, not just qualia.

Robinson, H. 1993. The anti-materialist strategy and the "knowledge argument". In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

Shoemaker, S. 1984. Churchland on reduction, qualia, and introspection. *Philosophy of Science Association* 1984, 2:799-809.

Introspection reveals functional properties, not physical, so qualia should be reduced to the functional, not to the physical.

Stemmer, N. 1989. Physicalism and the argument from knowledge. *Australasian Journal of Philosophy* 67:84-91.

Physicalism explains all the relevant evidence, hence all facts, and needn't admit mental entities; belief in mental entities is based on physical facts.

Thompson, E. 1992. Novel colors. *Philosophical Studies* 68:321-49.

Interesting remarks on what it would be for someone to see colors that we cannot, combining philosophical considerations with empirical findings about color space. Argues that science could tell us what such colors are like.

Tye, M. 2000. Knowing what it is like: The ability hypothesis and the knowledge argument. In *Consciousness, Color, and Content*. MIT Press.

Warner, R. 1986. A challenge to physicalism. *Australasian Journal of Philosophy* 64:249-65.

A Jackson-like argument that physical knowledge can't give you the knowledge of what pain feels like. With detailed consideration of objections and replies. Argues from limited incorrigibility to factualism about pains.

Watkins, M. 1989. The knowledge argument against the knowledge argument. *Analysis*, 49:158-60.

Epiphenomenalism => qualia don't cause beliefs => we don't know about qualia.

Zemach, E. 1990. Churchland, introspection, and dualism. *Philosophia* 20:3-13.

1.3b Zombies & Modal Arguments [see also 1.2c, 1.3a, 1.3c, 1.3d,

1.7e]

Balog, K. 1999. Conceivability, possibility, and the mind-body problem. *Philosophical Review* 108:497-528.

Bringsjord, S. 1999. The zombie attack on the computational conception of mind. *Philosophy and Phenomenological Research* 59:41-69.

Brueckner, A. 2001. Chalmers' conceivability argument for dualism. *Analysis*.

Byrne, A. 1999. Cosmic hermeneutics. *Philosophical Perspectives*.

Chalmers, D. J. 1996. Naturalistic dualism. In *The Conscious Mind*. Oxford University Press.

Argues from the lack of logical supervenience to the falsity of physicalism. A two-dimensional analysis shows that objections from a posteriori necessity fail. Argues for a naturalistic variety of property dualism.

Chalmers, D. J. 2002. Does conceivability entail possibility? In (T. Gendler & J. Hawthorne, eds) *Imagination, Conceivability, and Possibility*. Oxford University Press.

Cottrell, A. 1999. Sniffing the camembert: On the conceivability of zombies. *Journal of Consciousness Studies* 6:4-12.

Dennett, D. C. 1995. The unimagined preposterousness of zombies. *Journal of Consciousness Studies* 2:322-26.

Guzeldere, G. 1995. Varieties of zombiehood. *Journal of Consciousness Studies* 2:326-33.

Hill, C. S. 1997. Imaginability, conceivability, possibility, and the mind-body problem. *Philosophical Studies* 87:61-85.

Argues that the conceivability of zombies and the like can be explained away, in terms of the cognitive separability of perceptual imagination and sympathetic imagination of the same states.

Hill, C. S. 1998. Chalmers on the apriority of modal knowledge. *Analysis* 58:20-26.

Hill, C. S. & McLaughlin, B. P. 1998. There are fewer things in reality than are dreamt of in Chalmers' philosophy. *Philosophy and Phenomenological Research*.

Kirk, R. 1974. Sentience and behaviour. *Mind* 81:43-60.

Describing a situation where we would be justified in believing in zombies. Argues that zombies are

logically possible, which seems incompatible with most or all varieties of materialism.

Kirk, R. 1974. Zombies vs materialists. *Aristotelian Society Supplement* 48:135-52.

Materialism requires that physical states logically entail all non-relational states; but zombies are logically possible, so materialism fails. With a description of a zombie, and replies to a verificationist. All very true.

Kirk, R. 1977. Reply to Don Locke on zombies and materialism. *Mind* 86:262-4.

Reply to Locke 1976: materialism needs zombies to be logically impossible.

Kirk, R. 1999. Why there couldn't be zombies. *Proceedings of the Aristotelian Society, Supplementary Volume* 73:1-16.

Kraemer, E. R. 1980. Imitation-man and the 'new' epiphenomenalism. *Canadian Journal of Philosophy* 10:479-487.

If Campbell's imitation man is possible, then the causal relation between the physical and phenomenal is unreliable.

Latham, N. 1998. Chalmers on the addition of consciousness to the physical world. *Philosophical Studies*.

Levine, J. 1998. Conceivability and the metaphysics of mind. *Nous* 32:449-480.

Locke, D. 1976. Zombies, schizophrenics, and purely physical objects. *Mind* 83:97-99.

Contra Kirk: the logical possibility of zombies is compatible with empirical materialism. With some comments on Kirk's thought-experiment.

Marton, P. 1998. Zombies vs. materialists: The battle over conceivability. *Southwest Philosophy Review* 14:131-38.

Melnyk, A. 1998. Physicalism unfalsified: Chalmers' inconclusive argument for dualism. In (B. Loewer & C. Gillett, eds) *Physicalism and its Discontents*. Oxford University Press.

Moody, T. 1994. Conversations with zombies. *Journal of Consciousness Studies* 1:196-200.

Argues that behavioral differences in zombies would show up, in their discourse about consciousness.

Nagel, T. 1998. Conceiving the impossible and the mind-body problem. *Philosophy* 73:337-52.

Perkins, M. 1970. Matter, sensation, and understanding. *American Philosophical Quarterly* 8:1-12.

On the possibility of an Insentient Perceiver, who perceives the world without sensation. Sensation is inessential to perception and understanding, except understanding in the "whatlike" manner.

Perkins, M. 1971. Sentience. *Journal of Philosophy* 68:329-37.

Argues for the conceivability of insentient perception of colors (in "Insent", a kind of blindsighter or zombie), in order to argue for a realistic account of colors.

Perry, J. 2001. The zombie argument. In *Knowledge, Possibility, and Consciousness*. MIT Press.

Perry, J. 2001. The modal argument. In *Knowledge, Possibility, and Consciousness*. MIT Press.

Prudovsky, G. 1995. Arguments from conceivability. *Ratio* 8:63-69.

Robb, D. 1999. Conceivability and consciousness. *Philosophical Topics*.

Robinson, H. 1976. The mind-body problem in contemporary philosophy. *Zygon* 11:346-360.

A discussion of materialism and its difficulties. The conceivability of zombies poses special problems. Criticism of Smart's & Armstrong's analyses.

Squires, R. 1974. Zombies vs materialists II. *Aristotelian Society Supplement* 48:153-63.

Stalnaker, R. 2002. What is it like to be a zombie? In (T. Gendler & J. Hawthorne, eds) *Imagination, Coceivability, and Possibility*. Oxford University Press.

Stoljar, D. 2000. Physicalism and the necessary a posteriori. *Journal of Philosophy* 87:33-55.

Thomas, N. 1998. Zombie killer. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

van Gulick, R. 1999. Conceiving beyond our means: The limits of thought experiments. In (S. Hameroff, A. Kaszniak, & D. Chalmers, eds) *Toward a Science of Consciousness III*. MIT Press.

Yablo, S. 1998. Concepts and consciousness. *Philosophy and Phenomenological Research*.

Yablo, S. 1998. Textbook Kripkeanism and the open texture of language. *Philosophical Quarterly*.

1.3c Essentialist Arguments (Kripke) [see also 1.3b, 1.3d]

Kripke, S. A. 1971. Identity and necessity. In (M. Munitz, ed) *Identity and Individuation*.

An identity between mental and physical states can't be contingent, as it relates rigid designators. But nevertheless the co-occurrence of certain mental and physical states is contingent, so the identity theory is false.

Kripke, S. A. 1972. *Naming and Necessity*. Harvard University Press.

Both "pain" and "C-fibres firing" are rigid designators, so if they are identical, this must be necessary. But their co-occurrence is contingent, and this can't be explained away epistemically, so the identity theory fails.

Barnette, R. 1977. Kripke's pains. *Southern Journal of Philosophy* 15.

Argues that pain and the associated epistemic situation are inequivalent. Beliefs about pain are simply produced by mechanisms, and could come about without any sensation.

Bayne, S. R. 19xx. Kripke's Cartesian argument. *Philosophia*.

Trying to turn Kripke's argument against him: it's possible that pains and C-fibre stimulations are identical, so it's necessary that they're identical.

Bealer, G. 1994. Mental properties. *Journal of Philosophy* 91:185-208.

On four arguments against the identity theory: multiple-realizability, modal, knowledge, and certainty arguments. All face difficulties due to scientific essentialism, but the latter two can be reformulated to avoid them.

Blumenfeld, J. 1975. Kripke's refutation of materialism. *Australasian Journal of Philosophy* 53:151-6.

Kripke's argument doesn't refute token identity. Pains can have other essential properties besides painfulness, so psychophysical token identities can be necessary.

Boyd, R. 1980. Materialism without reductionism: What physicalism does not entail. In (N. Block, ed) *Readings in the Philosophy of Psychology*, vol 1. Harvard University Press.

Materialism doesn't need rigid identities, due to the compositional plasticity of mental states. So the possibility of disembodiment is compatible with materialism. The possibility of zombies is illusory.

Carney, J. & von Bretzel, P. 1973. Modern materialism and essentialism. *Australasian Journal of Philosophy* 51:78-81.

A materialist must deny essentialism to meet Kripke's argument.

Carney, J. 1975. Kripke and materialism. *Philosophical Studies* 27:279-282.

Comments on Feldman 1974: Feldman's view requires rejection of Kripke's views on necessity, or a problematic mixed view on rigid designators.

Della Rocca, M. 1993. Kripke's essentialist arguments against the identity theory. *Philosophical Studies* 69:101-112.

Kripke's premise that pains are essentially mental either begs the question (by assuming pains don't have physical properties) or weakens the premise that physical events aren't essentially mental.

Double, R. 1976. The inconclusiveness of Kripke's argument against the identity theory. *Auslegung* 3:156-65.

Feldman, F. 1973. Kripke's argument against materialism. *Philosophical Studies* 24:416-19.

Painfulness need not be an essential feature of pains.

Feldman, F. 1974. Kripke on the identity theory. *Journal of Philosophy* 71:665-76.

Kripke's arguments against person-body and mind-brain identity rely on the essentialness of aliveness to persons and painfulness to pains. There's no reason to grant this. If we do, rigidity is irrelevant to the argument.

Feldman, F. 1980. Identity, necessity, and events. In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol. 1. Harvard University Press.

Defending a contingent event identity thesis against Kripke. Mental properties (which are distinct from physical properties) may not be essential properties of an event.

Gjelsvik, O. 1988. A Kripkean objection to Kripke's arguments against the identity-theories. *Inquiry* 30:435-50.

Uses Kripke's 1979 direct-reference theory against him. When rigid designators don't have associated reference-fixing descriptions, we can't expect the "explaining away" strategy to work.

Hill, C. S. 1981. Why Cartesian intuitions are compatible with the identity thesis. *Philosophy and Phenomenological Research* 42:254-65.

The apparent contingency of identity is due to the fact that one can be aware of pain without being aware of C-fibers and vice versa, as well as to the fact that "C-fibers" may be picked out by a contingent description.

Holman, E. 1988. Qualia, Kripkean arguments, and subjectivity. *Philosophy Research Archives* 13:411-29.

Defending Kripkean arguments against various objections. Analysis in terms of manifest properties and their role in fixing reference to the subjective and objective.

Jackson, F. 1980. A note on physicalism and heat. *Australasian Journal of Philosophy* 58:26-34.

A Kripkean argument against non-analytic physicalism. Even if pain rigidly designates a brain state, the physicalist still has problems explaining the property of "pain-presents".

Jacquette, D. 1987. Kripke and the mind-body problem. *Dialectica* 41:293-300.

Kripke's argument doesn't refute contingent identity between minds and nonrigidly designated bodies, which is all materialism needs.

Leplin, J. 1979. Theoretical identification and the mind-body problem. *Philosophia* 8:673-88.

Some theoretical identification are analogous to mental-physical identifications -- entities are introduced by properties considered essential within a theory, but this doesn't preclude identification.

Levin, M. 1975. Kripke's argument against the identity thesis. *Journal of Philosophy* 72:149-67.

The reference of "pain" is fixed not by essential features but by contingent topic-neutral descriptions; this is the real moral of Wittgenstein's private language argument. So Kripke's apparent contingency can be explained away?

Levin, M. 1995. Tortuous dualism. *Journal of Philosophy* 92:313-22.

Reply to Bealer 1994. Tries to clarify the dialactic, and argues that the materialist can explain "possibility" of straw thought as thought conjoined with mere appearance of straw.

Lycan, W. G. 1974. Kripke and the materialists. *Journal of Philosophy* 71:677-89.

Kripke equivocates on "pain-sensation": pains aren't the same as impressions of pain. Argues that imaginability arguments aren't decisive, and that functionalism may be less vulnerable than the identity theory.

Lycan, W. G. 1987. Functionalism and essence. In *Consciousness*. MIT Press.

Painfulness needn't be essential to pains: pains are events, not objects, and events don't have essences; and the reference of "pain" is fixed by topic-neutral descriptions. With remarks on pains vs. pain-sensations.

Maxwell, G. 1979. Rigid designators and mind-brain identity. *Minnesota Studies in the Philosophy of Science* 9.

"Brain state" reference is fixed by topic-neutral description; picks out pain but mightn't have, explaining

the illusion of contingency. "Nonphysicalist materialism" results, with mind the essence of matter. An important paper.

McGinn, C. 1977. Anomalous monism and Kripke's Cartesian intuitions. *Analysis* 2:78-80. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Token identity theories aren't vulnerable to Kripke's argument: it may be essential to this pain that it is a C-fibre firing, although not to pain as a type.

McMullen, C. 1984. An argument against the identity theory. *Pacific Philosophical Quarterly* 65:277-87.

We can explain away the apparent contingency of identity in terms of possible differences in *evidence* for the physical state. With a discussion on identities between something perceived and something described.

Mucciolo, L. 1975. On Kripke's argument against the identity thesis. *Philosophia* 5:499-506.

"Pain" need not be a rigid designator, but instead may pick out a state by its causal role. If it is a rigid designator, then the apparent contingency of identity comes from imagining something else filling the causal role.

Sher, G. 1977. Kripke, Cartesian intuitions, and materialism. *Canadian Journal of Philosophy* 7:227-38.

The reference of "C-fibre stimulation" might be fixed contingently, allowing the intuitive contingency of identity to be explained away.

Taylor, P. 1983. McGinn, token physicalism, and a rejoinder of Woodfield. *Analysis* 43:80-83.

Woodfield, A. 1978. Identity theories and the argument from epistemic counterparts. *Analysis* 38:140-3.

Contra McGinn 1977, the counterpart strategy fails as any pain that occurred here now would have been this pain. A counterpart strategy on brain states may work. With a reply by McGinn and a later rejoinder by Woodfield.

1.3d Arguments from Disembodiment [see also 1.3b, 1.3c]

Alston, W. P. & Smythe, T. W. 1994. Swinburne's argument for dualism. *Faith and Philosophy* 11:127-33.

Carrier, L. 1974. Definitions and disembodied minds. *Personalist Forum* 55:334-43.

Cole, D. J. & Foelber, F. 1984. Contingent materialism. *Pacific Philosophical Quarterly* 65:74-85.

Argues that materialism is only contingently true, as it's conceptually possible that we could become immaterial by gradual replacement.

Hart, W. D. 1988. *The Engines of the Soul*. Cambridge University Press.

Lewy, C. 1943. Is the notion of disembodied existence self-contradictory? *Proceedings of the Aristotelian Society* 43:59-78.

Long, D. 1977. Disembodied existence, physicalism, and the mind-body problem. *Philosophical Studies*.

Merricks, T. 1994. A new objection to a priori arguments for dualism. *American Philosophical Quarterly* 31:81-85.

Physicalism is compatible with the possibility of disembodiment: one can hold that mind and body are identical, that the body is physical, but that it is not essentially physical.

Odegard, D. 1970. Disembodied existence and central state materialism. *Australasian Journal of Philosophy* 48:256-60.

Pecnjak, D. 1995. Remarks on disembodied existence. *Acta Analytica* 10:209-13.

Shoemaker, S. 19xx On an argument for dualism. Reprinted in *Identity, Cause, and Mind*. Cambridge University Press, 1984.

Swinburne, R. 1997. The modal argument for substance dualism. In *The Evolution of the Soul* (revised edition). Oxford University Press.

Taliaferro, C. 1986. A modal argument for dualism. *Southern Journal of Philosophy* 24:95-108.

Taliaferro, C. 1997. Possibilities in the philosophy of mind. *Philosophy and Phenomenological Research* 57:127-37.

Tidman, P. 1994. Conceivability as a test for possibility. *American Philosophical Quarterly* 31:297-309.

Tye, M. 1983. On the possibility of disembodied existence. *Australasian Journal of Philosophy* 61:275-282.

There's no reason to believe that disembodied existence is possible: lack of logical contradiction doesn't imply possibility, conceivability is too weak a criterion, and it's not obvious that the situation is imaginable.

van Cleve, J. 1983. Conceivability and the Cartesian argument for dualism. *Pacific Philosophical*

Quarterly.

Yablo, S. 1993. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53:1-42.

Zimmerman, D. 1991. Two Cartesian arguments for the simplicity of the soul. *American Philosophical Quarterly* 28:127-37.

1.3e Consciousness and Physicalism, Misc [see also 1.2, 1.3, 1.7b, 3.5b]

Fox, M. 1978. Beyond materialism. *Dialogue* 17:367-70.

Hill, C. S. 1991. *Sensations: A Defense of Type Materialism*. Cambridge University Press.

Defending type materialism, by way of criticism of dualism and functionalism. With treatments of introspection, sensory concepts, and other minds.

Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*. Holder-Pichler-Tempsky.

On why materialism requires conceptual analysis to locate mental properties in the natural world. Even a posteriori necessary connections have to be backed by a priori links. With remarks on supervenience. A nice paper.

Kirk, R. 1979. From physical explicability to full-blooded materialism. *Philosophical Quarterly* 29:229-37.

If every physical events has a physical explanation, and the mental is causally efficacious, then mental facts are strictly implied by physical facts. A nice argument.

Kirk, R. 1982. Physicalism, identity, and strict implication. *Ratio* 24:131-41.

Physicalism requires that all mental facts be strictly implied by the physical facts. Once this is recognized, questions about necessary or contingent identity are beside the point, and indeed identity is irrelevant.

Locke, D. 1971. Must a materialist pretend he's anaesthetized? *Philosophical Quarterly* 21:217-31.

Lund, D. 2000. Materialism and the subject of consciousness. *Idealistic Studies* 30:7-23.

Madell, G. 1988. *Mind and Materialism*. Edinburgh University Press.

On the problems posed for materialism by intentionality, autonomy, awareness, and indexicality.
Tentatively advocates a Cartesian position.

Robinson, H. 1982. *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Robinson, H. (ed) 1993. *Objections to Physicalism*. Oxford University Press.

A collection of arguments against physicalism, mostly based on worries about consciousness and qualia.

Robinson, W. S. 1982. Sellarsian materialism. *Philosophy of Science* 49:212-27.

Seager, W. E. 1992. *Metaphysics of Consciousness*. Routledge and Kegan Paul.

Consciousness could be physical even if not explicable; but supervenience worries make it hard to see how it *could* be physical, though causal role suggests that it must be. We need a new conception. A stimulating book.

Sellars, W. 1981. Is consciousness physical? *Monist* 64:66-90.

On the place of "occurrent pink" and the "sensorium" in the physical world. It may turn out that the physics of the brain differs from other physics, in order to accommodate the causal role of sensations.

Smith, A. D. 1993. Non-reductive physicalism? In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

A careful discussion of how to characterize physicalism, in terms of identity or supervenience, and argues that physicalism must reduce (bowdlerize) qualia to something they are not, as physicalism requires topic-neutral analyses.

1.3f Consciousness and Dualism [see also 1.2c, 1.3, 3.4d]

Collins, C. 1997. Searle on consciousness and dualism. *International Journal of Philosophical Studies* 5:15-33.

Double, R. 1983. Nagel's argument that mental properties are nonphysical. *Philosophy Research Archives* 9:217-22.

Eccles, J. 1987. Brain and mind: Two or one? In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Foster, J. 1989. A defense of dualism. In (J. Smythies & J. Beloff, eds) *The Case for Dualism*. University of Virginia Press.

Argues that all forms of materialism fail, and that dualism is the only option. Defends dualism against objections, and argues for interactionism over epiphenomenalism. A very clear and interesting paper.

Honderich, T. 1981. Psychophysical law-like connections and their problems. *Inquiry* 24:277-303.

Defending lawlike connections between physical states & conscious occurrents. Contra anomalous monism and identity theory for occurrents. But occurrents may not be causally efficacious. Comments by Wilson/Sprigge/Mackie/Stich.

Lahav, R. & Shanks, N. 1992. How to be a scientifically respectable 'property dualist'. *Journal of Mind and Behavior* 13:211-32.

Argues that the hypothesis of consciousness as an irreducible global property of the brain is compatible with what we know of both neuroscience and physics. With interesting remarks on quantum mechanics.

Latham, N. 2000. Chalmers on the addition of consciousness to the physical world. *Philosophical Studies* 98:67-93.

McGinn, C. 1993. Consciousness and cosmology: Hyperdualism ventilated. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

A dialogue with a "hyperdualist". On the pros and cons of materialist vs. dualist ontology and cosmology. Dualism avoids the "magic" of emergence at the cost of an inflated and bizarre ontology.

O'Leary-Hawthorne, J. & McDonough, J. K. 1998. Numbers, minds, and bodies: A fresh look at mind-body dualism. *Philosophical Perspectives* 12:349-71.

Rosenberg, J. F. 1988. On not knowing what or who one is: Reflections on the intelligibility of dualism. *Topoi* 7:57-63.

Smook, R. 1988. Egoicity and twins. *Dialogue* 27:277-86.

Smythies, J. R. & Beloff, J. (eds) 1989. *The Case for Dualism*. University of Virginia Press.

Sprigge, T. L. S. 1994. Consciousness. *Synthese* 98:73-93.

On the non-physical nature of consciousness, and the threat of a merely contingent connection to behavior; suggests a denial of "Hume's principle". Perhaps consciousness is the noumenal essence of the physical. A nice paper.

Swinburne, R. 1986. *The Evolution of the Soul*. Oxford University Press.

Taliaferro, C. 1996. *Consciousness and the Mind of God*. Cambridge University Press.

von Wright, G. H. 1994. On mind and matter. *Journal of Theoretical Biology* 171:101-10.

1.3g Mind-Body Problem, General

Beck, L. W. 1940. The psychophysical as a pseudo-problem. *Journal of Philosophy* 37:561-71.

Butler, C. W. 1972. The mind-body problem: A nonmaterialistic identity thesis. *Idealistic Studies* 2:229-48.

Campbell, K. K. 1970. *Body and Mind*. Doubleday.

Carrier, M. & Mittelstrass, J. 1991. *Mind, Brain, Behavior: The Mind-Body Problem and the Philosophy of Psychology*. de Gruyter.

Cheng, C. (ed) 1975. *Philosophical Aspects of the Mind-Body Problem*. Hawaii University Press.

Cooper, W. E. 1977. Beyond materialism and back again. *Dialogue* 16:191-206.

Diaz, J. 2000. Mind-body unity, dual aspect, and the emergence of consciousness. *Philosophical Psychology* 13:393-403.

Feigl, H. 1934. Logical analysis of the psychophysical problem. *Philosophy of Science* 1:420-45.

Feinberg, T. E. 1997. The irreducible perspectives of consciousness. *Seminars in Neurology* 17:85-93.

Fodor, J. A. 1981. The mind-body problem. *Scientific American* 244:114-25.

An overview: behaviorism, identity theory, functionalism, etc.

Foss, J. E. 1987. Is the mind-body problem empirical? *Canadian Journal of Philosophy* 17:505-32.

Yes it is. Empirical evidence bears on materialism, property dualism, emergentism, functionalism, interactive dualism, idealism, etc.

Gomes, G. 1995. Self-awareness and the mind-brain problem. *Philosophical Psychology* 8:155-65.

Gunderson, K. 1970. Asymmetries and mind-body perplexities. *Minnesota Studies in the Philosophy of Science* 4:273-309.

The core of the mind-body problem is the first/third-person asymmetry. It's like a periscope trying to place itself between its crosshairs. But this doesn't imply any strong ontological consequences.

Honderich, T. 1989. *Mind and Brain*. Oxford University Press.

Honderich, T. 1995. Consciousness, neural functionalism, and real subjectivity. *American Philosophical Quarterly* 32:369-381.

Against "neural functionalism", and on how Searle's view reduces to either neural functionalism or property dualism.

Howard, D. J. 1986. The new mentalism. *International Philosophical Quarterly* 26:353-7.

Hutto, D. D. 1998. An ideal solution to the problems of consciousness. *Journal of Consciousness Studies* 5:328-43.

Kim, J. 1997. The mind-body problem: Taking stock after forty years. *Philosophical Perspectives* 11:185-207.

Kneale, M. 1950. What is the mind-body problem? *Proceedings of the Aristotelian Society* 50:105-22.

Kohler, W. 1960. The mind-body problem. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

Levin, M. 1979. *Metaphysics and the Mind-Body Problem*. Oxford University Press.

Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford University Press.

On the mind-body problem and physical theory. Against reductive physicalism; instead, experience is the intrinsic nature of the physical. Discusses the interpretation of quantum mechanics and much else. A very enjoyable book.

Lowe, E. J. 1996. *Subjects of Experience*. Cambridge University Press.

Lund, D. H. 1994. *Perception, Mind, and Personal Identity: A Critique of Materialism*. University Press of America.

Margolis, J. 1974. Reductionism and ontological aspects of consciousness. *Journal for the Theory of Social Behavior* 4:3-16.

Matson, W. I. 1966. Why isn't the mind-body problem ancient? In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of

Minnesota Press.

Matson, W. I. 1976. *Sentience*. University of California Press.

McMullen, T. 1997. Sperry on consciousness as an emergent causal agent. *Australian Journal of Psychology* 49:152-155.

Nagel, T. 1993. What is the mind-body problem? In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

On ways in which we might locate consciousness within the natural world via scientific study. Perhaps we need an wider conception of objective reality.

Nagel, T. 1994. Consciousness and objective reality. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Nagel, T. 2001. The psychophysical nexus. In (P. Boghossian & C. Peacocke, eds) *New Essays on the A Priori*. Oxford University Press.

O'Shaughnessy, B. 1994. The mind-body problem. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Pratt, J. B. 1936. The present status of the mind-body problem. *Philosophical Review* 65:144-56.

Reber, A. 1997. Caterpillars and consciousness. *Philosophical Psychology* 10:437-49.

Ripley, C. 1984. Sperry's concept of consciousness. *Inquiry* 27:399-423.

An in-depth analysis of Sperry's views on consciousness. Sperry is not a dualist; he simply believes in "structural causation" based on emergent properties. Thorough and interesting.

Robinson, W. S. 1988. *Brains and People: An Essay on Mentality and its Causal Conditions*. Temple University Press.

Rosenthal, D. M. (ed) 1971. *Materialism and the Mind-Body Problem*. Prentice-Hall.

A collection of essays from the 1960s on the identity theory, functionalism, eliminative materialism.

Senchuk, D. M. 1991. Consciousness naturalized: Supervenience without physical determinism. *American Philosophical Quarterly* 28:37-47.

Sellars, W. 1953. A semantical solution of the mind-body problem. *Methodos* 5:45-84. Reprinted in *Pure*

Pragmatics and Possible Worlds. Ridgview, 1980.

Sellars, W. 1971. The double knowledge approach to the mind-body problem. *New Scholasticism* 45:269-89.

Silberstein, M. 1998. Emergence and the mind-body problem. *Journal of Consciousness Studies* 5:464-82.

Sperry, R. W. 1969. A modified concept of consciousness. *Psychological Review* 76:532-36.

Consciousness is an emergent property of brain dynamics that itself governs low-level flow of excitation.
Midway between mentalism and materialism.

Sperry, R. W. 1980. Mind-brain interaction: Mentalism yes, dualism no. *Neuroscience* 5:195-206.

A summary of the position whereupon mental properties are emergent and have independent causal powers. With a contrast to Popper and Eccles' dualism.

Sperry, R. W. 1992. Turnabout on consciousness: A mentalist view. *Journal of Mind and Behavior* 13:259-80.

An account of the "new mentalist paradigm". Clarifies earlier work, comments on others' interpretations. The view is monist and functionalist, but consciousness is a distinct emergent quality with a "downward" causal role.

Strawson, G. 1994. The experiential and the non-experiential. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Velmans, M. 1990. Consciousness, brain, and the physical world. *Philosophical Psychology* 3:77-99.

Wagner, S. J. 1994. Supervenience, recognition, and consciousness. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Warner, R. & Szubka, T. 1994. *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

A collection of 27 (mostly original) papers on the mind-body problem.

Weintraub, R. 1999. The spatiality of the mental and the mind-body problem. *Synthese* 117:409-17.

Wilson, D. L. 1976. On the nature of consciousness and of physical reality. *Perspectives in Biology and Medicine* 19:568-581.

Wisdom, J. 1957. Some main mind-body problems. *Proceedings of the Aristotelian Society* 60:187-210.

1.4

Specific Views on Consciousness [see also 1.2, 1.3, 1.5c]

1.4a Higher-Order Thought Approaches (Rosenthal, etc) [see also 6.2i]

Aquila, R. 1990. Consciousness as higher-order thoughts: Two objections. *American Philosophical Quarterly* 27:81-87.

Higher-order thought theories have two unacceptable consequences: one can notice one's hearing a sound without noticing one's consciousness of the sound; and one can unconsciously perceive one's surroundings as gloomy.

Byrne, A. 1997. Some like it HOT: consciousness and higher-order thoughts. *Philosophical Studies* 2:103-29.

Carruthers, P. 1989. Brute experience. *Journal of Philosophy* 258-69.

Argues for a distinction between conscious and non-conscious experiences, depending on whether one is conscious of the experience. Animal experiences are of the second kind, and therefore are not morally significant.

Carruthers, P. 1992. Consciousness and concepts. *Aristotelian Society Supplement* 66:41-59.

Advocates the "reflexive thinking" account of consciousness over Kirk's "presence" account. Availability for reflexive thinking is naturally necessary and sufficient for qualia. Interesting paper.

Carruthers, P. 1996. *Language, Thought, and Consciousness*. Cambridge University Press.

Carruthers, P. 1997. Fragmentary versus reflexive consciousness. *Mind and Language* 12:181-95.

Carruthers, P. 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.

Dretske, F. 1993. Conscious experience. *Mind* 102:263-283.

Against higher-order thought accounts: one can have a conscious experience without being aware that one is having it. With remarks on thing-awareness vs. fact-awareness and on "inner-sense" accounts.

Dretske, F. 1995. Are experiences conscious? In *Naturalizing the Mind*. MIT Press.

We're not conscious *of* our experience in general, but conscious *with* it. Criticizes HOP theories (not conceptualized enough) and HOT theories (rules out animals; there's more in experience than thought).

Francescotti, R. M. 1995. Higher-order thoughts and conscious experience. *Philosophical Psychology*.

Argues that a higher-order thought is insufficient for consciousness, even with Rosenthal's constraint. A causal constraint is required, but the only strong enough such constraint doesn't work.

Gennaro, R. J. 1993. Brute experience and the higher-order thought theory of consciousness. *Philosophical Papers* 22:51-69.

Carruthers 1989 misanalyzes higher-order thought theory. There's no need for conscious HOTs, and not too much conceptual sophistication is required, so animals might have HOTs and therefore conscious pains.

Gennaro, R. J. 1996. *Consciousness and Self-consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*. John Benjamins.

Guzeldere, G. 1996. Consciousness and the introspective link principle. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Guzeldere, G. 1995. Is consciousness the perception of what passes in one's own mind? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

A critique of higher-order-perception theories of consciousness. They're either committed to a "representational divide" fallacy or collapse into higher-order-thought or first-order theories.

Jamieson, D. & Bekoff, M. 1992. Carruthers on nonconscious experience. *Analysis* 52:23-28.

Various points against Carruthers 1989. His examples of nonconscious experience are likely conscious, and the higher-order account is circular.

Jacob, P. 1996. State consciousness revisited. *Acta Analytica* 11:29-54.

Kobes, B. W. 1995. Telic higher-order thoughts and Moore's paradox. *Philosophical Perspectives* 9:291-312.

Levine, J. 1997. Are qualia just representations? (Critical notice of Tye.) *Mind and Language* 12:101-13.

Lycan, W. G. 1995. Consciousness as internal monitoring, I. *Philosophical Perspectives* 9:1-14.

Argues for a Lockean quasi-perceptual view of consciousness as internal monitoring via second-order states. Contra objections, e.g. Rey's point that it makes consciousness too prevalent -- consciousness isn't

an on-off affair.

Lycan, W. 2001. A simple argument for a higher-order representation theory of consciousness. *Analysis* 61:3-4.

Mellor, D. H. 1978. Conscious belief. *Proceedings of the Aristotelian Society* 78:87-101.

Conscious belief (or assent) is believing that one believes. Addresses various objections, from self-deception and from consciousness of assent. Communication needs conscious belief, not just belief.

Mellor, D. H. 1980. Consciousness and degrees of belief. In (D. H. Mellor, ed) *Prospects for Pragmatism*. Cambridge University Press.

Natsoulas, T. 1992. Appendage theory -- pro and con. *Journal of Mind and Behavior* 13:371-96.

On various pros and cons of HOT theories, to do with reflexivity, objects of HOTs, introspection, and so on. With comparisons to "intrinsic" theories.

Natsoulas, T. 1992. Are all instances of phenomenal experience conscious in the sense of their being objects of inner (second-order) consciousness? *American Journal of Psychology* 105:605-12.

Natsoulas, T. 1993. What is wrong with the appendage theory of consciousness? *Philosophical Psychology* 6:137-54.

On three theories of our direct awareness of conscious states: mental-eye theories, self-intimational theories, and appendage theory. Appendage theory (i.e. HOT theory) is promising, but how does an HOT determine its object?

Natsoulas, T. 1993. The importance of being conscious. *Journal of Mind and Behavior* 14:317-40.

On the differences between first-order and second-order consciousness. Second-order consciousness is essential for communication and locomotion. With remarks on "nonconscious consciousness".

Nelkin, N. 1989. Unconscious sensations. *Philosophical Psychology* 2:129-41.

Separates CN (phenomenological) from C1 (info-processing) and C2 (higher order beliefs). CN is a subset of CS (image-representation state). We are always C2 of CN states, but not of other CS states: unconscious sensations!

Nelkin, N. 1995. The dissociation of phenomenal states from apperception. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Argues that we need not be apperceptively aware of phenomenal states. Introspection leaves the matter open, but some empirical results (e.g. hue blindsight) and theoretical arguments support dissociability.

Ridge, M. 2001. Taking solipsism seriously: Nonhuman animals and meta-cognitive theories of consciousness. *Philosophical Studies* 103:315-340.

Rosenthal, D. M. 1986. Two concepts of consciousness. *Philosophical Studies* 49:329-59.

Consciousness should be construed neither as sensation nor intentionality, but as the existence of higher-order thoughts.

Rosenthal, D. M. 1997. A theory of consciousness. In (N. Block, O. Flanagan, and G. Guzeldere, eds) *The Nature of Consciousness*. MIT Press.

A conscious mental state is a state that is the subject of a higher-order thought. Consciousness is not essential to mentality, should be separated from sensory quality, and is not an intrinsic property of conscious states.

Rosenthal, D. M. 1990. Why are verbally expressed thoughts conscious? Bielefeld Report.

Because verbally expressing and reporting are easily and immediately connected for 1st-order thoughts. But not for 2nd-order thoughts. Hmmm.

Rosenthal, D. M. 1993. Thinking that one thinks. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Conscious states are states that are the contents of higher-order thoughts. Express/report distinction: we report them, and express the HOT (which may be unconscious). Defense against dispositional and collapsing objections.

Rosenthal, D. M. 1993. Explaining consciousness. Manuscript.

Distinguishes the sense in which we are aware of conscious states; argues for the separation of consciousness and sensation; and outlines how higher-order thoughts might explain the what-it's-like of conscious states.

Rosenthal, D. M. 1993. Higher-order thoughts and the appendage theory of consciousness. *Philosophical Psychology* 6:155-66.

In response to Natsoulas, HOT theory needn't answer the general question of how intentional states determine their objects. With remarks on the other alternatives and the dangers of self-intimation.

Rosenthal, D. M. 1995. Moore's paradox and consciousness. *Philosophical Perspectives* 9:313-33.

Rosenthal, D. M. 1997. Apperception, sensation, and dissociability. *Mind and Language* 2:206-23.

- Rosenthal, D. M. 1998. Consciousness and metacognition. In (D. Sperber, ed) *Metarepresentation*. Oxford University Press.
- Rosenthal, D. M. 2000. Consciousness, interpretation, and consciousness. *Protosociology* 14.
- Rowlands, M. 2001. Consciousness and higher-order thoughts. *Mind and Language* 16:290-310.
- Seager, W. E. 1994. Dretske on HOT theories of consciousness. *Analysis* 54:270-76.
- Seager, W. E. 1999. HOT Theory: The mentalistic reduction of consciousness. In *Theories of Consciousness: An Introduction and Assessment*. Routledge.
- Stamenov, M. I. 1997. Grammar, meaning, and consciousness: What sentence structure can tell us about the structure of consciousness. In (M. Stamenov, ed) *Language Structure, Discourse, and the Access to Consciousness*. John Benjamins.
- Stoerig, P. 1997. Phenomenal vision and apperception: Evidence from blindsight. *Mind and Language* 2:224-37.
- Stone, J. 2001. What is it like to have an unconscious mental state? *Philosophical Studies* 104:179-202.
- Thomasson, A. 2000. After Brentano: A one-level theory of consciousness. *European Journal of Philosophy* 8:190-210.

1.4b Dennett on Consciousness [see also 1.7c]

- Akins, K. 1996. Lost the plot? Reconstructing Dennett's multiple drafts theory of consciousness. *Mind and Language* 11:1-43.
- Akins, K, & Winger, S. 1996. Ships in the night: Churchland and Ramachandran on Dennett's theory of consciousness. In (K. Akins, ed) *Perception*. Oxford University Press.
- Arbib, M. A. 1972. Consciousness: The secondary role of language. *Journal of Philosophy* 69.
- Baker, L. R. 1995. Content meets consciousness. *Philosophical Topics* 22:1-22.
- Block, N. 1995. What is Dennett's theory a theory of? *Philosophical Topics* 22:23-40.
- Bricke, J. 1984. Dennett's eliminative arguments. *Philosophical Studies* 45:413-29.

Criticizing Dennett's accounts of pains, dreams, and images: in no case do his arguments earn their

eliminative conclusions.

Bricke, J. 1985. Consciousness and Dennett's intentionalist net. *Philosophical Studies* 48:249-56.

Reportability is no good for capturing consciousness: it completely leaves out the qualitative content of conscious states.

Churchland, P. S. & Ramachandran, V. S. 1993. Filling in: Why Dennett is wrong. In (B. Dahlbom, ed) *Dennett and His Critics*. Blackwell.

Argues that Dennett's account of the blindspot and scotomas are wrong. Neurophysiological data suggests that blind areas are represented explicitly; psychological data shows that it's not just "more of the same".

Clark, S. R. L. 1993. Minds, memes, and rhetoric. *Inquiry* 36:3-16.

Dennett, D. C. 1968. *Content and Consciousness*. Routledge.

Dennett, D. C. 1978. Reply to Arbib and Gunderson. In *Brainstorms*. MIT Press.

On various notions of awareness: contents of the speech center, contents directing behavior, and contents of attention. We have privileged access to one sort, but it is a different sort that plays the main role in control.

Dennett, D. C. 1978. Toward a cognitive theory of consciousness. *Minnesota Studies in the Philosophy of Science*, Vol. 9. Reprinted in *Brainstorms* (MIT Press, 1978).

Conscious contents are contents of a buffer memory to which a public report module has access. We only have conscious access to propositional judgments, not to underlying processes. With a cute functional diagram.

Dennett, D. C. 1979. On the absence of phenomenology. In (D. Gustafson & B. Tapscott, eds) *Body, Mind, and Method*. Kluwer.

There is no real phenomenology. There are only *judgments* about phenomenology, and nothing more is going on. We don't have privileged access to anything, except perhaps certain propositional episodes.

Dennett, D. C. 1982. How to study human consciousness empirically, or, nothing comes to mind. *Synthese* 53:159-80.

We can study consciousness by the method of heterophenomenology: studying the things we say about conscious states, which we can interpret as we interpret texts. Autophenomenology gives nothing extra. With comments by Rorty.

Dennett, D. C. 1988. The evolution of consciousness. Manuscript.

Consciousness is a virtual machine which evolved.

Dennett, D. C. 1991. *Consciousness Explained*. Little-Brown.

Argues against the "Cartesian Theatre", advocating a "multiple drafts" model of consciousness. Presents a detailed model of processes underlying verbal report, and argues that there is nothing else (e.g. qualia) to explain.

Dennett, D. C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15:183-201.

Using temporal anomalies in consciousness to support a "Multiple Drafts" theory of consciousness rather than a "Cartesian Theater". Contents of consciousness are wholly determined by effects on action/memory.

Dennett, D. C. 1993. *Precis of Consciousness Explained*. *Philosophy and Phenomenological Research* 53:889-931.

A discussion of *Consciousness Explained*, with comments by Tye, Jackson, Shoemaker, and Rosenthal, and a reply by Dennett.

Dennett, D. C. 1993. Living on the edge. *Inquiry* 36:135-59.

A reply to Clark, Fellows & O'Hear, Foster, Lockwood, Seager, Siewert, and Sprigge.

Dennett, D. C. 1993. Caveat emptor. *Consciousness and Cognition* 2:48-57.

A reply to Baars & McGovern, Mangan, Toribio.

Dennett, D. C. 1995. Is perception the "leading edge" of memory. In (A. Spafadora, ed) *Memory and Oblivion*.

There is no "leading edge" of consciousness, separating perception and memory. With an analysis of metacontrast cases, etc.

Dennett, D. C. 1995. Get real. *Philosophical Topics* 22:505

Dennett, D. C. 1996. Seeing is believing -- or is it? In (K. Akins, ed) *Perception*. Oxford University Press.

Dennett, D. 2001. Are we explaining consciousness yet? *Cognition* 79:221-37.

Densmore, S. & Dennett, D. C. 1999. The virtues of virtual machines. *Philosophy and Phenomenological Research* 59:747-61.

Dretske, F. 1995. Differences that make no difference. *Philosophical Topics* 22:41-57.

Criticizes Dennett's first-person operationalism as Cartesian. There can be awareness without judgment -- e.g. non-epistemic perception. This comes from information or "micro-judgments", and is not conceptual.

Fellows, R. & O'Hear, A. 1993. Consciousness avoided. *Inquiry* 36: 73-91.

Foster, J. 1993. Dennett's rejection of dualism. *Inquiry* 36:17-31.

Gunderson, K. 1972. *Content and Consciousness* and the mind-body problem. *Journal of Philosophy* 69.

Hutto, D. 1995. Consciousness demystified: A Wittgensteinian critique of Dennett. *Monist* 78:464-79.

Jackson, F. 1993. Appendix A (for philosophers). *Philosophy and Phenomenological Research* 53:897-901.

Presses Dennett on the "truth-maker" question for materialists: what sort of physical fact makes it true that people are conscious?

Johnsen, B. 1997. Dennett on qualia and consciousness: A critique. *Canadian Journal of Philosophy* 27:47-82.

Kirk, R. 1993. "The best set of tools"? Dennett's metaphors and the mind-body problem. *Philosophical Quarterly* 43:335-43.

Joycean machines and multiple drafts turn out to shed no light on the question of what features make a conscious system conscious.

Lockwood, M. 1993. Dennett's mind. *Inquiry*.

Argues for a suitably sophisticated Cartesian Theatre, and against the identification of phenomenology with judgments.

Mangan, B. 1993. Dennett, consciousness, and the sorrows of functionalism. *Consciousness and Cognition* 2:1-17.

Marbach, E. 1988. How to study consciousness phenomenologically or quite a lot comes to mind. *Journal of the British Society for Phenomenology*, 19:252-268.

Marbach, E. 1994. Troubles with heterophenomenology. In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*. Holder-Pichler-Tempsky.

McCauley, R. N. 1993. Why the blind can't lead the blind: Dennett on the blind spot, blindsight, and sensory qualia. *Consciousness and Cognition* 2:155-64.

Brings empirical evidence to bear against Dennett's "filling-in" account of the blindspot, and argues that blindsight and the blindspot aren't analogous.

McGinn, C. 1995. Consciousness evaded: Comments on Dennett. *Philosophical Perspectives* 9:241-49.

Nikolinakos, D. 2000. Dennett on qualia: The case of pain, smell and taste. *Philosophical Psychology* 13:505-522.

Puccetti, R. 1993. Dennett on the split-brain. *Psychology* 4(52).

Radner, D. 1994. Heterophenomenology: Learning about the birds and the bees. *Journal of Philosophy* 91:389-403.

Rey, G. 1995. Dennett's unrealistic psychology. *Philosophical Topics* 22:259-89.

Robinson, W. S. 1972. Dennett's analysis of awareness. *Philosophical Studies* 23:147-52.

Robinson, W. S. 1994. Orwell, Stalin, and determinate qualia. *Pacific Philosophical Quarterly* 75:151-64.

Dennett's Orwell/Stalin argument doesn't establish its conclusion, as "brain smear" is quite compatible with determinate qualia.

Rockwell, T. 1996. Awareness, mental phenomena, and consciousness: A synthesis of Dennett and Rosenthal. *Journal of Consciousness Studies* 3:463-76.

Rorty, R. 1972. Dennett on awareness. *Philosophical Studies* 23:153-62.

Rosenthal, D. M. 1993. Multiple drafts and higher-order thoughts. *Philosophy and Phenomenological Research* 53:911-18.

Rosenthal, D. M. 1994. First-person operationalism and mental taxonomy. *Philosophical Topics* 22:319-349.

Rosenthal, D. M. 1995. Multiple drafts and the facts of the matter. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

We can subtract first-person operationalism from Dennett's multiple-drafts account, giving a higher-order thought theory.

Rosenthal, D. 2000. Content, interpretation, and consciousness. *Protosociology* 14:67-84.

Seager, W. E. 1993. Verification, skepticism, and consciousness. *Inquiry*.

An elucidation of Dennett's fundamental eliminativism about phenomenology, resting on verificationist arguments. Like many sceptical arguments, it ends up too powerful to be convincing.

Seager, W. E. 1999. Dennett, part I and II. In *Theories of Consciousness: An Introduction and Assessment*. Routledge.

Shoemaker, S. 1993. Lovely and suspect ideas. *Philosophy and Phenomenological Research* 53:903-908.

Siewert, C. 1993. What Dennett can't imagine and why. *Inquiry*.

Argues that zombies are conceivable, via partial zombiehood in blindsight patients who respond unprompted. Dennett's arguments rely on a question-begging third-person absolutism.

Sprigge, T. L. S. 1993. Is Dennett a disillusioned zimbo? *Inquiry* 36:33-57.

Toribio, J. 1993. Why there still has to be a theory of consciousness. *Consciousness and Cognition* 2:28-47.

Criticizes behavioral, localist, and "intransitive" approaches to consciousness, and recommends a "transitive" metacognitive approach. But criticizes Dennett for not explaining subjective experience.

Tye, M. 1993. Reflections on Dennett and consciousness. *Philosophy and Phenomenological Research* 53:891-6.

Argues that Dennett's verificationism begs the question, and that "seeming" cannot be identified with believing or judging.

Van Gulick, R. 1995. Dennett, drafts, and phenomenal realism. *Philosophical Topics* 22:443-55.

Wuketits, F. 1994. Consciousness explained -- or explained away? *Acta Analytica* 9:55-64.

1.4c Functionalism [see also 1.4a, 1.4b, 1.5c, 1.8, 3.4]

Levin, J. 1991. Analytic functionalism and the reduction of phenomenal states. *Philosophical Studies* 61:211-38.

Contra Kripkean arguments, a good enough functional theory may help close the conceivability/explanatory gap between the physical and qualia. Contra Nagel/Jackson, such a theory could provide us with recognitional abilities.

Mangan, B. 1998. Against functionalism: Consciousness as an information-bearing medium. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Marcel, A. 1988. Phenomenal experience and functionalism. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Myin, E. 1998. Holism, functionalism and visual awareness. *Communication and Cognition*, 31:3-19.

Perlis, D. 1995. Consciousness and complexity: The cognitive quest. *Annals of Mathematics and Artificial Intelligence* 14:309-21.

Shoemaker, S. 1993. Functionalism and consciousness. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

Argues that introspective access is essential to many sorts of mental state, due to constitutive rationality requirements. Against a perceptual model of introspection; introspecting and introspected states are closer than that.

Schweizer, P. 1996. Physicalism, functionalism, and conscious thought. *Minds and Machines* 6:61-87.

van Gulick, R. 1988. A functionalist plea for self-consciousness. *Philosophical Review* 97:149-88.

How functionalism can handle consciousness: Self-consciousness is the possession of reflexive metapsychological information. This helps understand learning, representation, and belief. Phenomenal experience is still tricky.

1.4d Eliminativism [see also 1.7c, 2.1c, 3.5c]

Allport, A. 1988. What concept of consciousness? In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Churchland, P. S. 1983. Consciousness: the transmutation of a concept. *Pacific Philosophical Quarterly* 64:80-95.

Experimental evidence against consciousness/introspection/transparency.

Dennett, D. C. 1976. Are dreams experiences? *Philosophical Review* 73:151-71. Reprinted in *Brainstorms* (MIT Press, 1978).

Argues that dreams might not be experienced, but rather be stored directly into memory (the "cassette-tape" theory of dreaming).

Dennett, D. C. 1979. The onus re experiences: A reply to Emmett. *Philosophical Studies* 35, 315-18.

Emmett, K. 1978. Oneiric experiences. *Philosophical Studies* 34:445-50.

Nikolinakos, D. 1994. General anesthesia, consciousness, and the skeptical challenge. *Journal of Philosophy* 2:88-104.

Consciousness is an indispensable concept in anesthesiology, and therefore (contra Churchland and Wilkes) is a scientifically legitimate kind. With empirical details and anesthesiological theory on levels of consciousness.

Rey, G. 1982. A reason for doubting the existence of consciousness. In (R. Davidson, S. Schwartz, & D. Shapiro, eds) *Consciousness and Self-Regulation*, Vol 3. Plenum Press.

One could make a machine, duplicating the usual abilities that go along with consciousness, but surely it wouldn't be conscious. So what are the conditions for consciousness? Maybe there are none.

Rey, G. 1986. A question about consciousness. In (H. Otto & J. Tuedio, eds) *Perspectives on Mind*. Kluwer.

A rerun of Rey 1982: An unconscious machine could duplicate all the obvious criteria for consciousness, so maybe even we aren't conscious. With remarks on the relation between our belief in consciousness and consciousness itself.

Rey, G. 1995. Toward a projectivist account of conscious experience. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

We "project" consciousness into ourselves and others. There are no explanation-transcendent phenomena for which there is non-question-begging evidence. With remarks on self-attribution and Wittgenstein.

Smith, D. W. 1986. In (H. Otto & J. Tuedio, eds) *Perspectives on Mind*. Kluwer.

Commentary on Rey 1986: we are directly aware of our consciousness. It's not a theoretical entity, but rather something to be explained.

Tienson, J. L. 1987. Brains are not conscious. *Philosophical Papers* 16:187-93.

A skeptical argument: single neurons are not conscious, and adding a neuron won't produce consciousness, so finite brains are not conscious.

Wilkes, K. V. 1984. Is consciousness important? *British Journal for the Philosophy of Science* 35:223-43.

No, and it's not very coherent either. It divides into awakeness, sensation, sensory experience, and

propositional attitudes. Also a history of the term.

Wilkes, K. V. 1988. Yishi, Duh, Um and consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Wilkes, K. V. 1995. Losing consciousness. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Consciousness is not a tenable notion in either commonsense or scientific psychology; we should return instead to the "psuche".

Williams, D. C. 1934. Scientific method and the existence of consciousness. *Psychological Review* 41:461-79.

Williams, D. C. 1959. Mind as a matter of fact. *Review of Metaphysics* 13:205-25.

1.4e Epiphenomenalism [see also 1.3a, 1.3f, 1.6e, 3.6]

Bieri, P. 1992. Trying out epiphenomenalism. *Erkenntnis* 36:283-309.

Birnbacher, D. 1988. Epiphenomenalism as a solution to the ontological mind-body problem. *Ratio* 1:17-32.

Chalmers, D. J. 1996. The paradox of phenomenal judgment. In *The Conscious Mind*. Oxford University Press.

Considers major arguments against the causal or explanatory irrelevance of consciousness -- arguments from self-knowledge, memory, reference, etc -- and argues that none pose fatal flaws.

Creel, R. 1980. Radical epiphenomenalism: B. F. Skinner's account of private events. *Behaviorism* 8:31-53.

Dennett, D. C. 1991. "Epiphenomenal" qualia? In *Consciousness Explained*, pp. 398-406. Little-Brown.

Discusses two senses of "epiphenomenalism" -- "Huxley's" and "philosophical" varieties -- and argues that the philosophical sort is crazy. (N.B. Huxley actually subscribed to the "philosophical" variety.)

Double, R. 1979. Taylor's refutation of epiphenomenalism. *Journal of Critical Analysis* 8:23-28.

Flanagan, O. & Polger, G. 1998. Consciousness, adaptation, and epiphenomenalism. In (G. Mulhauser, ed) *Evolving Consciousness*.

Hodges, M. 1979. Meaning and the impotence hypothesis. *Review of Metaphysics* 32:515-29.

Horowitz, A. 1999. Is there a problem in physicalist epiphenomenalism? *Philosophy and Phenomenological Research* 59:421-34.

Huxley, T. 1874. On the hypothesis that animals are automata. *Fortnightly Review* 95:555-80. Reprinted in *Collected Essays*. London, 1893.

Hyslop, A. 1998. Methodological epiphenomenalism. *Australasian Journal of Philosophy* 78:61-70.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Uses the Mary thought-experiment to argue that qualia are epiphenomenal, and argues that epiphenomenalism is a tenable doctrine.

Kraemer, E. R. 1980. Imitation-man and the 'new' epiphenomenalism. *Canadian Journal of Philosophy* 10:479-487.

If Campbell's imitation man is possible, then the causal relation between the physical and phenomenal is unreliable.

Lachs, J. 1963. Epiphenomenalism and the notion of cause. *Journal of Philosophy* 60:141-45.

Lachs, J. 1963. The impotent mind. *Review of Metaphysics* 17:187-99.

Lachs, J. 1967. Angel, animal, machine: Models for man. *Southern Journal of Philosophy* 5:221-27.

Long, W. 1953. Comments on the alleged proof of epiphenomenalism. *British Journal for the Philosophy of Science* 3:355-58.

Pauen, M. 2000. Painless pain: Property dualism and the causal role of phenomenal consciousness. *American Philosophical Quarterly* 37:51-64.

Pecnjak, D. 1989. Epiphenomenalism and machines: A discussion of van Rooijen's critique of Popper. *British Journal for the Philosophy of Science* 40:404-8.

Popper, K. R. 1977. Some remarks on panpsychism and epiphenomenalism. *Dialectica* 31:177-86.

Puccetti, R. 1974. Physicalism and the evolution of consciousness. *Canadian Journal of Philosophy Supplement* 1:171-83.

Robinson, D. 1993. Epiphenomenalism, laws, and properties. *Philosophical Studies* 69:1-34.

Argues that phenomenal information implies epiphenomenalism, even at the intra-psychic level. With remarks on ineffability and on whether properties should be individuated by nomic role or by essence.

Robinson, W. S. 1982. Causation, sensation, and knowledge. *Mind* 91:524-40.

Argues that we can have non-inferential knowledge of sensations even if they don't make a causal difference to our beliefs, so epiphenomenalism is OK. All theories have a similar problems.

Rudd, A. 2000. Phenomenal judgment and mental causation. *Journal of Consciousness Studies* 7:53-69.

van Rooijen, J. 1987. Interactionism and evolution: A critique of Popper. *British Journal for the Philosophy of Science* 38:87-92.

Wasserman, G. D. 1982. Materialism and mentality. *Review of Metaphysics* 35:715-30.

Wassermann, G. 1979. Reply to Popper's attack on epiphenomenalism. *Mind* 88:572-75.

Watkins, M. 1989. The knowledge argument against the knowledge argument. *Analysis*, 49:158-60.

Epiphenomenalism => qualia don't cause beliefs => we don't know about qualia.

Wisdom, J. O. 1954. Is epiphenomenalism refutable? *Proceedings of the 2nd International Congress of the International Union for the Philosophy of Science* 5:73-78.

Woodhouse, M. 1974. A new epiphenomenalism? *Australasian Journal of Philosophy* 52:163-69.

1.4f Interactionism [see also 1.3f, 3.3d]

Beloff, J. 1994. Minds and machines: A radical dualist perspective. *Journal of Consciousness Studies* 1:32-37.

Buncombe, M. 1995. *The Substance of Consciousness: An Argument for Interactionism*. Avebury.

Elitzur, A. C. 1989. Consciousness and the incompleteness of the physical explanation of behavior. *Journal of Mind and Behavior* 10:1-20.

Argues from the fact that we talk about consciousness to the conclusion that consciousness plays an active role, so physical laws must be incomplete.

Elitzur, A. C. 1990. Neither idealism nor materialism: A reply to Snyder. *Journal of Mind and Behavior*.

- Elitzur, A. C. 1995. Consciousness can no longer be ignored. *Journal of Consciousness Studies* 2:353-58.
- Foster, J. 1991. *The Immaterial Self: A Defense of the Cartesian Dualist Conception of Mind*. Routledge.
- Hodgson, D. 1991. *The Mind Matters: Consciousness and Choice in a Quantum World*. Oxford University Press.
- Jackson, F. 1980. Interactionism revived? *Philosophy of Social Science* 10:316-23.
- Libet, B. 1994. A testable theory of mind-brain interaction. *Journal of Consciousness Studies* 1:119-26.
- Lindahl, B. I. B. & Arhem, P. 1996. Mind as a force field: Comments on a new interactionistic hypothesis. *Journal of Theoretical Biology* 171:111-22.
- Popper, K. & Eccles, J. 1977. *The Self and Its Brain: An Argument for Interactionism*. Springer.
- Popper, K. 1994. *Knowledge and the Body-Mind Problem: In Defence of Interaction*. Routledge.
- Roelofs, H. D. 1955. A case for dualism and interactionism. *Philosophy and Phenomenological Research* 15:451-76.
- Snyder, D. 1990. On Elitzur's discussion of the impact of consciousness on the physical world. *Journal of Mind and Behavior*.

Argues with Elitzur on quantum mechanics and consciousness. With response.

- Warner, R. 1996. Facing ourselves: Incorrigibility and the mind-body problem. *Journal of Consciousness Studies* 3:217-30.
- Wilson, D. L. 1999. Mind-brain interactionism and the violation of physical laws. *Journal of Consciousness Studies*.

1.4g Panpsychism [see also 1.4h]

- Bjelland, A. G. 1982. Popper's critique of panpsychism and process proto-mentalism. *Modern Schoolman* 59:233-43.
- Butler, C. 1978. Panpsychism: A restatement of the genetic argument. *Idealist Studies* 8:33-39.
- Chalmers, D. J. 1996. Is experience ubiquitous? In *The Conscious Mind*. Oxford University Press.

There are no strong arguments against panpsychism, and good reason to take it seriously. Extrapolating the processing properties crucial for standard complex experience suggests that simple process may yield simple experience.

de Quincey, C. 1994. Consciousness all the way down? *Journal of Consciousness Studies* 1:217-29.

An analysis of a debate between Griffin and McGinn on panexperientialism, arguing for new forms of understanding.

Edwards, P. 1967. Panpsychism. In (P. Edwards, ed) *The Encyclopedia of Philosophy*, volume 5. Macmillan.

An excellent review article on panpsychism; highly recommended.

Farleigh, P. 1998. Whitehead's even more dangerous idea. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Ford, M. P. 1981. William James: Panpsychist and metaphysical realist. *Transactions of the Peirce Society* 17:158-70.

Griffin, D. R. 1997. Panexperiential physicalism and the mind-body problem. *Journal of Consciousness Studies* 4:248-68.

An interesting paper arguing for an experiential aspect in all matter, explicating a Whiteheadian position.

Griffin, D. R. 1998. *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem*. University of California Press.

Griffin, D. R. 1998. Pantemporalism and panexperientialism. In (P. Harris, ed) *The Textures of Time*. University of Michigan Press.

Hartshorne, C. 1978. Panpsychism: Mind as sole reality. *Ultim Real Mean* 1:115-29.

Hut, P. & Shepard, R. 1996. Turning the "hard problem" upside-down and sideways. *Journal of Consciousness Studies* 3:313-29.

Argues for a new fundamental feature ("X") which stands to consciousness as time stands to motion, thus making consciousness possible and ubiquitous.

Nagel, T. 1979. Panpsychism. In *Mortal Questions*. Cambridge University Press.

Material composition, nonreductionism, realism, non-emergence => panpsychism.

Popper, K. R. 1977. Some remarks on panpsychism and epiphenomenalism. *Dialectica* 31:177-86.

Rensch, B. 1977. Argument for panpsychist identism. In (J. Cobb & D. Griffin, eds) *Mind in Nature*. University Press of America.

Robinson, E. A. 1949. Animism as a world hypothesis. *Philosophical Review* 58:53-63.

Rosenberg, G. H. 1996. Rethinking nature: A hard problem within the hard problem. *Journal of Consciousness Studies* 3:76-88.

On why consciousness extends beyond the cognitive. Argues that fundamental laws for consciousness must connect at a basic level, and argues that panpsychism is not as implausible as often thought.

Seager, W. 1995. Consciousness, information, and panpsychism. *Journal of Consciousness Studies* 2:272-88. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem* (MIT Press, 1999).

Examines a position on which experience is fundamental to the world, and suggests that this ought to lead to panpsychism. With some connections to information and quantum mechanics.

Sellars, R. W. 1960. Panpsychism or evolutionary materialism. *Philosophy of Science* 27:329-49.

Shepherd, J. J. 1974. Panpsychism and parsimony. *Process Studies* 4:3-10.

Sprigge, T. L. S. 1983. The vindication of panpsychism. In *The Vindication of Absolute Idealism*. Edinburgh University Press.

van Cleve, J. 1990. Mind -- dust or magic? Panpsychism versus emergence. *Philosophical Perspectives* 4:215-226.

On Nagel 1979: emergence is more plausible than panpsychism. A construal of emergence as nomological supervenience without logical supervenience.

Wright, S. 1953. Gene and organism. *American Naturalist*.

Wright, S. 1977. Panpsychism and science. In (J. Cobb & D. Griffin, eds) *Mind in Nature*. University Press of America.

1.4h Intrinsic Monism (Russell, etc) [see also 1.4h]

Blackburn, S. 1992. Filling in space. *Analysis* 52:62-3.

Physics is dispositional, but if there are only bare dispositions, then the world has no nature of its own.

And if there are categorical grounds, we have no idea what they could be, except maybe subjective qualia.

Chalmers, D. J. 1996. The metaphysics of information. In *The Conscious Mind*, pp. 301-8 (see also pp. 153-55). Oxford University Press.

An "it from bit" view fits the Russellian metaphysics (described earlier): physics is info from the outside, (proto)experience is info from the inside. The problem is constituting macrophenomenal from microphenomenal; some ideas.

Demopolous, W. & Friedman, M. 1989. The concept of structure in Russell's *The Analysis of Matter*. In (C. Savage & C. Anderson, eds) *Rereading Russell: Essays in Bertrand Russell's Metaphysics and Epistemology*. University of Minnesota Press.

A nice account of Russell's (and Schlick's and Carnap's) structuralism and Newman's objection, with analysis. (N.B. no philosophy of mind.)

Feigl, H. 1958. The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science* 2:370-497. Reprinted (with a postscript) as *The 'Mental' and the 'Physical'*. University of Minnesota Press, 1967.

A long and very interesting essay on the mind-body problem. Ultimately advocates a "structural" view of the physical and identifies experience with the underlying reality, at least for some neurophysiological states.

Feigl, H. 1960. The mind-body problem: Not a pseudo-problem. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

Feigl, H. 1971. Some crucial issues of mind-body monism. *Synthese* 22:295-312.

Feigl, H. 1975. Russell and Schlick: A remarkable agreement on a monistic solution of the mind-body problem. *Erkenntnis* 9:11-34.

Argues that Russell's and Schlick's views on the structural nature of physics and the possible identification of the "content" with experience are quite close to each other. With interesting historical remarks.

Feser, E. 1998. Can phenomenal qualities exist unperceived? *Journal of Consciousness Studies* 4:405-14.

Foster, J. 1982. *The Case for Idealism*. Routledge.

Foster, J. 1991. Lockwood's hypothesis. In *The Immaterial Self*. Oxford University Press.

Argues that the Russellian view is implausible, as the structure of the underlying physical processes does not correspond to the structure and quality of consciousness.

Lockwood, M. 1981. What was Russell's neutral monism? *Midwest Studies in Philosophy* 6:143-58.

Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford University Press.

On the mind-body problem and physical theory. Against reductive physicalism; instead, experience is the intrinsic nature of the physical. Explores some potential links with quantum mechanics.

Lockwood, M. 1993. The grain problem. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

On the "grain problem" for the intrinsic-nature view: how do lots of microphysical qualities add up into a smooth experience? Appeals to quantum mechanics and a preferred set of observables.

Lockwood, M. 1998. Unsensed phenomenal qualities: A defence. *Journal of Consciousness Studies* 4:415-18.

Maxwell, G. 1971. Structural realism and the meaning of theoretical terms. *Minnesota Studies in the Philosophy of Science*.

Maxwell, G. 1978. Unity of consciousness and mind-brain identity. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Maxwell, G. 1979. Rigid designators and mind-brain identity. *Minnesota Studies in the Philosophy of Science* 9.

"Brain state" reference is fixed by topic-neutral description; picks out pain but mightn't have, explaining the illusion of contingency. "Nonphysicalist materialism" results, with mind the essence of matter. An important paper.

Newman, M. H. A. 1928. Mr. Russell's causal theory of perception. *Mind*.

Argues against Russell's structuralism: any collection can be arranged to have a given structure, under some relation, so if physics tells us only about structure, it tells us at most the cardinality of the world.

Robinson, H. 1982. Matter: Turning the tables. In *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Rosenberg, G. H. 1997. *A Place for Consciousness: Probing the Deep Structure of the Natural World*. Dissertation, Indiana University.

Russell, B. 1927. *The Analysis of Matter*. London: Kegan Paul.

Argues that physics characterizes the external world only structurally, and leaves intrinsic qualities

unspecified. Only experience acquaints us with anything intrinsic. Perhaps the intrinsic nature of physics is experiential?

Schlick, M. 1925. *General Theory of Knowledge*.

Stoljar, D. 1997. Neutral monism. Manuscript.

Stoljar, D. 2001. Two conceptions of the physical. *Philosophy and Phenomenological Research* 62:253-81.

Stubenberg, L. 1996. The place of qualia in the world of science. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Stubenberg, L. 1997. Austria vs. Australia: Two versions of the identity theory. In (K. Lehrer & J. Marek, eds) *Austrian Philosophy, Past and Present*. Kluwer.

Stubenberg, L. 1998. *Consciousness and Qualia*. John Benjamins.

1.5 Consciousness and Content

1.5a

Consciousness and Intentionality (Searle, etc)

Cole, D. 1994. Thought and qualia. *Minds and Machines* 4:283-302.

Crane, T. 1998. Intentionality as the mark of the mental. *Philosophy*.

Davies, M. 1995. Consciousness and the varieties of aboutness. In (C. Macdonald, ed) *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford University Press.

Dunlop, C. 2000. Searle's unconscious mind. *Philosophical Psychology* 13:123-148.

Fodor, J. & Lepore, E. 1994. What is the Connection Principle? *Philosophy and Phenomenological Research* 54:837-45.

Searle's formulation of the connection principle is unclear, and there is no formulation is both plausible and interesting.

Georgalis, N. 1996. Awareness, understanding, and functionalism. *Erkenntnis* 44:225-56.

- Gillett, E. 1996. Searle and the "deep unconscious". *Philosophy, Psychiatry, and Psychology* 3:191-200.
- Gunderson, K. 1990. Consciousness and intentionality: Robots with and without the right stuff. In (C. A. Anderson & J. Owens, eds) *Propositional Attitudes: The Role of Content in Language, Logic, and Mind*. CSLI.
- Marbach, E. 1993. *Mental Representation and Consciousness: Toward a Phenomenological Theory of Representation and Reference*. Kluwer.
- Meijers, A. 2000. Mental causation and Searle's impossible conception of unconscious intentionality. *International Journal of Philosophical Studies* 8:155-170.
- McLoughlin, J. 1999. Unwittingly recapitulating Freud: Searle's concept of a vocabulary of the unconscious. *Ratio* 12:34-53.
- Natsoulas, T. 1992. Intentionality, consciousness, and subjectivity. *Journal of Mind and Behavior* 13:281-308.
- Nelkin, N. 1989. Propositional attitudes and consciousness. *Philosophy and Phenomenological Research* 49:413-30.

About conscious beliefs. We are not "conscious of" beliefs, merely "conscious that" -- i.e. belief is not phenomenological.

- Nelkin, N. 1993. The connection between intentionality and consciousness. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Against Searle: some intentional states aren't even potentially conscious (blindsight, etc) and intentional content doesn't require a particular phenomenal feel. So there's no essential link. With remarks on McGinn.

- Schweizer, P. 1994. Intentionality, qualia, and mind/brain identity. *Minds and Machines* 4:259-82.
- Seager, W. E. 1999. Conscious intentionality and the anti-Cartesian catastrophe. In *Theories of Consciousness: An Introduction and Assessment*. Routledge.
- Searle, J. R. 1984. Intentionality and its place in nature. *Synthese*.

(Subjective) intentionality sure is real. It causes and is caused.

- Searle, J. R. 1990. Consciousness, explanatory inversion and cognitive science. *Behavioral and Brain Sciences* 13:585-642.

Advocates a "connection principle": intentional states must be potentially conscious. If not, they're brutally neurophysiological. So cog-sci talk of "intentional" cognitive mechanisms below the conscious level isn't justified.

Searle J. R. 1994. The connection principle and the ontology of the unconscious: A reply to Fodor and Lepore. *Philosophy and Phenomenological Research* 54:847-55.

Clarifying the connection principle -- it's necessary in order to see how certain nonconscious neural states qualify as unconscious mental states.

van Gulick, R. 1988. Consciousness, intrinsic intentionality, and self-understanding machines. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

van Gulick, R. 1995. Why the connection argument doesn't work. *Philosophy and Phenomenological Research* 55:201-7.

van Gulick, R. 1995. How should we understand the relation between intentionality and phenomenal consciousness. *Philosophical Perspectives* 9:271-89.

Worley, S. 1997. Belief and consciousness. *Philosophical Psychology* 10:41-55.

Argues that belief requires consciousness, as we can't make sense of the personal/subpersonal content distinction without appealing to consciousness.

1.5b The Content of Experience

Baldwin, T. 1992. The projective theory of sensory content. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Bermudez, J. L. 1994. Peacocke's argument against the autonomy of nonconceptual representational content. *Mind and Language* 9:402-18.

Bermudez, J. 2000. Naturalized sense data. *Philosophy and Phenomenological Research* 61:353-374.

Bermudez, J. L. 1995. Nonconceptual content: From perceptual experience to subpersonal computational states. *Mind and Language* 10:333-69.

Berger, G. 1987. On the structure of visual sentience. *Synthese* 71:355-70.

Bilgrami, A. 1994. On McDowell on the content of perceptual experience. *Philosophical Quarterly* 44:206-13.

- Brewer, B. 1999. *Perception and Reason*. Oxford University Press.
- Butchvarov, P. 1980. Adverbial theories of consciousness. *Midwest Studies in Philosophy* 5:261-80.
- Church, J. 2000. 'Seeing as' and the double bind of consciousness. *Journal of Consciousness Studies* 7:99-112.
- Clark, R. 1973. Sensuous judgments. *Nous* 7:45-56.
- Clark, R. 1976. The sensuous content of perception. In (H. Castaneda, ed) *Action, Knowledge, and Reality*. Bobbs-Merrill.
- Clark, R. 1981. Sensing, perceiving, thinking. *Grazer Philosophische Studien* 12:273-95.
- Crane, T. (ed) 1992. *The Contents of Experience: Essays on Perception*. Cambridge University Press.
- Crane, T. 1992. The nonconceptual content of experience. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.
- Davies, W. M. 1996. *Experience and Content: Consequences of a Continuum Theory*. Avebury.
- DeBellis, M. 1991. The representational content of musical experience. *Philosophy and Phenomenological Research* 51:303-24.
- Contra Peacocke, we don't need sensational properties to understand the content of musical experience. Fine-grained representational properties can do the job, with the help of some Schenkerian analysis.
- Gunther, Y. H. 1995. Perceptual content and the subpersonal. *Conference* 6:31-45.
- Gunther, Y. H. 2001. Content, illusion, partition. *Philosophical Studies* 102:185-202.
- Hamlyn, D. W. 1994. Perception, sensation, and non-conceptual content. *Philosophical Quarterly* 44:139-53.
- Jackson, F. 1976. The existence of mental objects. *American Philosophical Quarterly* 13:33-40.
- Kelly, S. D. 2001. The non-conceptual content of perceptual experience: Situation dependence and fineness of grain. *Philosophy and Phenomenological Research* 62:601-608.
- Kraut, R. 1982. Sensory states and sensory objects. *Nous* 16:277-93.
- Langsam, H. 2000. Experiences, thoughts, and qualia. *Philosophical Studies* 99:269-295.

- Lowe, E. J. 1992. Experience and its objects. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.
- Martin, M. G. F. 1992. Perception, concepts, and memory *Philosophical Review* 101:745-63.
- McDowell, J. 1994. The content of perceptual experience. *Philosophical Quarterly* 44:190-205.
- McFarland, D. 1998. Crane on concepts and experiential content. *Analysis* 58:54-58.
- Millar, A. 1991. Concepts, experience, and inference. *Mind* 100:495-505.
- Natsoulas, T. 1983. What are the objects of perceptual consciousness? *American Journal of Psychology* 96:435-67.
- Natsoulas, T. 1994. On the distinction between the object and the content of consciousness. *Journal of Mind and Behavior* 15:239-64.
- Noe, A. 1999. Thought and experience. *American Philosophical Quarterly* 36:257-65.
- Peacocke, C. 1983. *Sense and Content: Experience, Thought, and their Relations*. Oxford University Press.
- Peacocke, C. 1984. Colour concepts and colour experience. *Synthese* 58:365-82.
- Peacocke, C. 1989. Perceptual content. In (J. Almog, J. Perry, & H. Wettstein, eds) *Themes from Kaplan*. Oxford University Press.
- Peacocke, C. 1992. Scenarios, concepts, and perception. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.
- Peacocke, C. 1994. Nonconceptual content: Kinds, rationales, and relations. *Mind and Language* 4:419-29.
- Peacocke, C. 1997. Nonconceptual content defended. *Philosophy and Phenomenological Research*.
- Peacocke, C. 2001. Does perception have a nonconceptual content? *Journal of Philosophy* 98:239-264.
- Peacocke, C. 2001. Phenomenology and nonconceptual content. *Philosophy and Phenomenological Research* 62:609-615.

Pendlebury, M. 1987. Perceptual representation. *Proceedings of the Aristotelian Society* 87:91-106.

Pendlebury, M. 1990. Sense experiences and their contents: A defense of the propositional account. *Inquiry* 33:215-30.

Lots of reasons why experiences have propositional content (i.e., their content is truth-evaluable, etc). A nice paper.

Schantz, R. 2001. The given regained: Reflections on the sensuous content of experience. *Philosophy and Phenomenological Research* 62:167-180.

Sedivy, S. 1996. Must conceptually informed perceptual experience involve nonconceptual content? *Canadian Journal of Philosophy* 26:413-31.

Snowdon, P. 1990. The objects of perceptual experience. *Aristotelian Society Supplement*, 64:121-50.

Sturgeon, S. 1998. Visual experience. *Proceedings of the Aristotelian Society* 72:179-200.

Tolhurst, W. 1998. Seemings. *American Philosophical Quarterly* 35:293-302.

Valberg, J. J. 1992. *The Puzzle of Experience*. Oxford University Press.

Yoon, B. 2000. Intentionality of perceptual experience. *Erkenntnis* 52:339-355.

1.5c

Representationalism

Beckermann, A. 1995. Visual information-processing and phenomenal consciousness. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Block, N. 1990. Inverted earth. *Philosophical Perspectives* 4:53-79.

Uses Inverted Earth case, colors and lenses inverted, to argue vs Harman that qualitative states aren't intentional states. Also, less convincingly, to argue that qualitative states aren't functional states.

Block, N. 1996. Mental paint and mental latex. In (E. Villanueva, ed) *Perception*. Ridgeview.

Block, N. 1999. Sexism, racism, ageism, and the nature of consciousness. *Philosophical Topics* 26.

Block, N. 2000. Mental paint. In (M. Hahn & M. Ramberg), *Essays on Burge*. MIT Press.

Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.

Harman, G. 1990. The intrinsic quality of experience. *Philosophical Perspectives*.

There are no real qualia problems, just Intentional confusions.

Harman, G. 1996. Explaining objective color in terms of subjective reactions. In (E. Villaneuva, ed) *Perception*. Ridgeview.

Lalor, B. 1999. Intentionality and qualia. *Synthese* 121:249-290.

Levine, J. 1997. Are qualia just representations? (Critical notice of Tye) *Mind and Language* 12:101-13.

Lloyd, D. 1991. Leaping to conclusions: connectionism, consciousness, and the computational mind. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Suggests that conscious states are identical to representational states, and that unconscious representation is impossible; transition between conscious states is non-representational. Appeals to connectionist models in support.

Lloyd, D. 1997. Consciousness and its discontents. *Communication and Cognition* 30:273-284.

Argues that consciousness and representation are distinct, as e.g. the latter depends on context but the former does not.

Lycan, W. G. 1996. Layered perceptual representation. In (E. Villaneuva, ed) *Perception*. Ridgeview.

Lycan, W. G. 1996. *Consciousness and Experience*. MIT Press.

Lycan, W. G. 1998. In defense of the representational theory of qualia. *Philosophical Perspectives* 12:479-87.

McCulloch, G. 1993. The very idea of the phenomenological. *Proceedings of the Aristotelian Society* 67:39-57.

The phenomenological can be reduced to the intentional. Intentional states have a what-it-is-like, and there is no special phenomenal object of introspection.

Neander, K. 1998. The division of phenomenal labor: A problem for representationalist theories of consciousness. *Philosophical Perspectives* 12:411-34.

O'Brien, G. & Opie, J. 1997. Cognitive science and phenomenal consciousness: A dilemma, and how to

avoid it. *Philosophical Psychology* 10:269-86.

Rey, G. 1998. A narrow representationalist account of qualitative experience. *Philosophical Perspectives* 12:435-58.

Robinson, W. S. 1998. Intrinsic qualities of experience: Surviving Harman's critique. *Erkenntnis* 47:285-309.

Seager, W. E. 1999. Representational theories of consciousness, parts I and II. In *Theories of Consciousness*. Routledge.

Shoemaker, S. 1990. Qualities and qualia: What's in the mind? *Philosophy and Phenomenological Research Supplement* 50:109-131.

Qualia can't be reduced to standard intentional properties (due to certain inversion cases). Projectivist and sense-reference accounts don't work either. Perhaps qualia are necessarily-illusory intentional properties.

Shoemaker, S. 1991. Qualia and consciousness. *Mind* 100:507-24.

On the relationship between phenomenal and intentional aspects of qualia, and in particular on the accessibility of qualia to conscious awareness. Phenomenal & intentional similarity are connected but must be distinguished.

Stalnaker, R. 1996. On a defense of the hegemony of representation. In (E. Villanueva, ed) *Perception*. Ridgeview.

Sullivan, P. R. 1995. Contentless consciousness and information-processing theories of mind. *Philosophy, Psychiatry, and Psychology* 2:51-59.

Tye, M. 1992. Visual qualia and visual content. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Tye, M. 1994. Do pains have representational content? In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*. Holder-Pichler-Tempsky.

Tye, M. 1995. A representational theory of pains and their phenomenal character. *Philosophical Perspectives* 9:223-39.

Argues that pain is representational, and that its phenomenal character is narrow nonconceptual content. They have a complex representational structure, with map-like arrays of sentential contents.

Tye, M. 1995. What "what it is like" is like. *Analysis*.

Argues that "what it is like to be X" is an intentional context, which solves some of the associated problems.

Tye, M. 1996. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press.

Tye, M. 1996. Orgasms again. In (E. Villanueva, ed) *Perception*. Ridgeview.

Tye, M. 1997. The problem of simple minds: Is there anything it's like to be a honeybee? *Philosophical Studies* 88:289-317.

Tye, M. 1998. Inverted earth, swampman, and representationalism. *Philosophical Perspectives* 12:459-78.

Stoljar, D. 1996. What what it's like isn't like. *Analysis* 56:281-83.

Vinueza, A. 2000. Sensations and the language of thought. *Philosophical Psychology* 13:373-392.

Wager, A. 1999. The extra qualia problem: Synaesthesia and representationism. *Philosophical Psychology* 12:263-281.

Warfield, T. 1999. Against representational theories of consciousness. *Journal of Consciousness Studies* 6:66-69.

White, S. 1994. Color and notional content. *Philosophical Topics* 22:471-503.

1.5d

Internalism and Externalism about Experience [see also 2.2]

Davies, M. 1992. Perceptual content and local supervenience. *Proceedings of the Aristotelian Society* 66:21-45.

Argues that perceptual content does not supervene on internal state, even though it is non-conceptual. Constructs an Twin scenario to that effect. With remarks on the relation between perceptual content and phenomenology.

Davies, M. 1993. Aims and claims of externalist arguments. In (E. Villanueva, ed) *Naturalism and Normativity*. Ridgeview.

Distinguishes modal and constitutive externalism, characterizes perceptual content and its relation to sensational content, and argues for externalism about perceptual content by examples.

de Vries, W. A. 1996. Experience and the swamp creature. *Philosophical Studies* 82:55-80.

Argues that a swampthing isn't intelligent or intentional, with different physiological processes and no sensations, as these are functional kinds.

Dretske, F. 1996. Phenomenal externalism, or if meanings ain't in the head, where are qualia? In (E. Villanueva, ed) *Perception*. Ridgeview.

We only have access to qualia through our concepts, which are external; so internal qualia would be inaccessible. So if qualia are knowable, they're external; and if not, why posit them? With comments by Kim, Horwich, Biro.

Forbes, G. 1997. Externalism and scientific Cartesianism. *Mind and Language* 12:196-205.

Kirk, R. 1994. The trouble with ultra-externalism. *Proceedings of the Aristotelian Society* 68:293-307.

Kirk, R. 1996. Why ultra-externalism goes too far. *Analysis* 56:73-79.

Kirk, R. 1998. Consciousness, information, and external relations. *Communication and Cognition* 30:249-71.

McCulloch, G. 1990. Externalism and experience. *Analysis* 50:244-50.

Argues against McGinn that one should embrace a form of "strong externalism" about experience. Experience can be laden with externally-grounded concepts.

McCulloch, G. 1994. Not much trouble for ultra-externalism. *Analysis* 54:265-9.

Sartwell, C. 1995. Radical externalism concerning experience. *Philosophical Studies* 78:55-70.

There is no epistemically available aspect of experience that is determined internally; experiences are "fused" with the environment.

Tappenden, P. 1996. The roundsquare copula: A semantic internalist's rejoinder. *Proceedings of the Aristotelian Society* 96:395-400.

1.5e Miscellaneous

Cam, P. 1984. Consciousness and content-formation. *Inquiry* 27:381-98.

Carruthers, P. 1998. Conscious thinking: Language or elimination? *Mind and Language* 13:457-476.

Falk, B. 1993. Consciousness, cognition, and the phenomenal. *Proceedings of the Aristotelian Society* 67:55-73.

On conceptual influences on experience, and aspectual seeing, focusing on bodily and dynamic elements. Self-awareness is not of phenomenal states but **in** them. With commentary by S. Mulhall.

Jacquette, D. 1984. Sensation and intentionality. *Philosophical Studies* 47:229-40.

Sensations don't have intentional objects, they *are* intentional objects.

Maloney, J. C. 1986. Sensuous content. *Philosophical Papers* 15:131-54.

McGinn, C. 1988. Consciousness and content. *Proceedings of the British Academy* 74:219-39. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Comparing the problems of consciousness and content, and reconciling optimism on content with pessimism on consciousness. The phenomenological nature of content may be mysterious, but the individuation of contents is not.

Nelkin, N. 1994. Phenomena and representation. *Philosophy of Science* 45:527-47.

Arguing against the view that phenomenal properties are "read off" in making perceptual judgments. Experiences do not literally have color or shape.

Sosa, E. 1986. Experience and intentionality. *Philosophical Topics* 14:67-83.

On a propositional conception of experience, and making sense of awareness of experience and various problems for sense-data monadicism.

1.6

Aspects of Consciousness

1.6a

Self-Consciousness [see also 6.2o]

Anscombe, G. E. M. 1975. The first person. In (S. Guttenplan, ed) *Mind and Language*. Oxford University Press.

Balaban, O. 1990. *Subject and Consciousness: A Philosophical Inquiry into Self-Consciousness*. Rowman & Littlefield.

- Bealer, G. 1997. Self-consciousness. *Philosophical Review* 106:69-117.
- Bermudez, J. 1997. Reduction and the self. *Journal of Consciousness Studies* 4:458-66.
- Bermudez, J. 1998. *The Paradox of Self-Consciousness*. MIT Press.
- Canfield, J. V. 1990. *The Looking-Glass Self: An Examination of Self-Awareness*. Praeger.
- Campbell, J. 1994. *Past, Space, and Self*. MIT Press.
- Campbell, J. 1995. The body image and self-consciousness. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Cassam, Q. 1995. Transcendental self-consciousness, In (P. Kumar, ed) *The Philosophy of P. F. Strawson*. Indian Council for Philosophical Research.
- Cassam, Q. 1997. *Self and World*. Oxford University Press.
- Castaneda, H. 1989. The reflexivity of self-consciousness: Sameness/identity, data for artificial intelligence. *Philosophical Topics* 17:27-58.
- Cheeks, J. M. & Briggs, S. R. 1982. Self-consciousness and aspects of personality. *Journal of Research in Personality* 16:401-8.
- Chisholm, R. M. 1969. On the observability of the self. *Philosophy and Phenomenological Research* 30:7-21.
- Christofidou, A. 2000. Self-consciousness and the double immunity. *Philosophy* 75:539-570.
- Church, J. 1990. Judgment, self-consciousness, and object-independence. *American Philosophical Quarterly* 27:51-60.
- Davis, L. H. 1989. Self-consciousness in chimps and pigeons. *Philosophical Psychology* 2:249-59.
- Delius, H. 1981. *Self-Awareness: A Semantical Inquiry*. Beck.
- Dennett, D. C. 1992. The self as the center of narrative gravity. In (F. Kessel, P. Cole, & D. L. Johnson, eds) *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.
- Eilan, N. 1995. Consciousness and the self. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and*

the Self. MIT Press.

Eilan, N. Marcel, A. J. & Bermudez, J. 1995. Self-consciousness and the body: Interdisciplinary issues. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.

Falk, A. 1995. Consciousness and self-reference. *Erkenntnis* 43:151-80.

Frith, U. & Happe, F. 1999. Theory of mind and self-consciousness: What is it like to be autistic? *Mind and Language* 14:1-22.

Gallagher, S. & Meltzoff, A. 1996. The earliest sense of self and others: Merleau-Ponty and recent developmental studies. *Philosophical Psychology* 9:211-33.

Gallagher, S. 1996. The moral significance of primitive self-consciousness: A response to Bermudez. *Ethics* 107:129-40.

Gallagher, S. 2000. Self-reference and schizophrenia: A cognitive model of immunity to error through misidentification. In (D. Zahavi, ed) *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience*. John Benjamins.

Ganeri, J. 2000. Cross-modality and the self. *Philosophy and Phenomenological Research* 61:639-657.

Gennaro, R. J. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333-47.

Argues that consciousness entails having episodic memory, which entails self-consciousness. So consciousness entails self-consciousness.

Gennaro, R. J. 1996. *Consciousness and Self-consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*. John Benjamins.

Hurley, S. L. 1998. Nonconceptual self-consciousness and agency: Perspective and access. *Communication and Cognition* 30:207-247.

James, W. 1890. The consciousness of self. In *The Principles of Psychology*.

Laycock, S. W. 1998. Consciousness it/self. *Journal of Consciousness Studies* 5:141-152.

Martin, M. G. F. 1995. Bodily awareness: A sense of ownership. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.

McCullagh, M. 2000. Functionalism and self-consciousness. *Mind and Language* 15:481-499.

- Meijsing, M. 2000. Self-consciousness and the body. *Journal Of Consciousness Studies* 7:34-50.
- Metzinger, T. 2000. The subjectivity of subjective experience: A representationist analysis of the first-person perspective. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Mittal, K. K. 1979. Self-identity and self-consciousness. *Indian Philosophical Quarterly* 7:159-63.
- Neisser, U. 1991. Five kinds of self-knowledge. *Philosophical Psychology* 1:35-59.
- O'Hear, A. 1989. Evolution, knowledge, and self-consciousness. *Inquiry* 32:127-150.
- Parker, S. T. , Mitchell, R. M. , & Boccia, M. L. 1994. *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.
- Proust, J. 2000. Awareness of agency: Three levels of analysis. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Richards, W. 1984. Self-consciousness and agency. *Synthese* 61:149-71.
- Self-consciousness is consciousness of agency. Castaneda/Nozick/Nagel.
- Shoemaker, S. 1968. Self-reference and self-awareness. *Journal of Philosophy* 65:555-67.
- Shoemaker, S. 1986. Introspection and the self. *Midwest Studies in Philosophy*.
- Shoemaker, S. 1994. Self-knowledge and "inner sense". *Philosophy and Phenomenological Research* 54:249-314.
- Smith, D. W. 1986. The structure of (self-) consciousness. *Topoi* 5:149-56.
- Sosa, E. 1983. Consciousness of self and of the present. In (J. Tomberlin, ed) *Agent, Language, and the Structure of the World*. Hackett.
- Stephens, G. L. & Graham, G. 1994. Self-consciousness, mental agency, and the clinical psychopathology of thought-insertion. *Philosophy, Psychiatry, and Psychology* 1:1-10.
- Stephens, G. L. & Graham, G. 2000. *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. MIT Press.
- Strawson, P. F. 1974. Self, mind, and body. In *Freedom and Resentment and Other Essays*.

Strawson, G. 1997. 'The self'. *Journal of Consciousness Studies* 4:405-28.

van Gulick, R. 1988. A functionalist plea for self-consciousness. *Philosophical Review* 97:149-88.

How functionalism can handle consciousness: Self-consciousness is the possession of reflexive metapsychological information. This helps understand learning, representation, and belief. Phenomenal experience is still tricky.

Varela, F. G. 1971. Self-consciousness: Adaptation or epiphenomenon? *Stud Gen* 24:426-39.

White, S. 1987. What is it like to be a homunculus? *Pac Philosophical Quarterly* 68:148-74.

Weird examples of homunculi that are conscious but not self-conscious. Self-consciousness, not consciousness, is what really counts.

Zahavi, D. 2000. Self and consciousness. In (D. Zahavi, ed) *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience*. John Benjamins.

1.6b The Unity of Consciousness [see also 5.12d, 6.1e, 6.1g]

Arvidson, P. 2000. Transformations in consciousness: Continuity, the self and marginal consciousness. *Journal Of Consciousness Studies* 7:3-26.

Bayne, T. 2000. The unity of consciousness: Clarification and defence. *Australasian Journal Of Philosophy* 78:248-254.

Beahrs, J. O. 1982. *Unity and Multiplicity: Multilevel Consciousness of Self in Hypnosis, Psychiatric Disorder, and Mental Health*. Brunner/Mazel.

Beahrs, J. O. 1983. Co-consciousness: A common denominator in hypnosis, multiple personality, and normalcy. *American Journal of Clinical Hypnosis* 26:100-13.

Brooks, D. H. M. 1985. Strawson, Hume, and the unity of consciousness. *Mind* 94:583-86.

Brooks, D. H. M. 1995. *The Unity of the Mind*. St. Martin's Press.

Cotterill, R. M. J. 1995. On the unity of conscious experience. *Journal of Consciousness Studies* 2:290-311.

Dainton, B. 2000. *Stream of Consciousness: Unity and continuity in conscious experience*. Routledge.

- Eccles, J. 1985. *The Brain and the Unity of Conscious Experience*. Cambridge University Press.
- Feinberg, T. 2000. The nested hierarchy of consciousness: A neurobiological solution to the problem of mental unity. *Neurocase* 6:75-81.
- Fox, I. 1985. The individualization of consciousness. *Philosophical Topics* 13:119-43.
- Hamlyn, D. W. 1996. The unity of the senses and self-consciousness. In *Understanding Perception: The Concept and its Conditions*. Avebury Press.
- Hill, C. S. 1991. Unity of consciousness, other minds, and phenomenal space. In *Sensations: A Defense of Type Materialism*. Cambridge University Press.
- Humphrey, N. 2000. One-self: A meditation on the unity of consciousness. *Social Research* 67:1059-1066.
- Hurley, S. 1994. Unity and objectivity. In (C. Peacocke, ed) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford University Press.
- Hurley, S. 1998. Unity, neuropsychology, and action. In *Consciousness in Action*. Harvard University Press.
- James, W. 1895. The knowing of things together. *Psychological Review* 2:105-24.
- Kobes, B. 2000. Unity of consciousness and bi-level externalism. *Mind and Language* 15:528-544.
- Lockwood, M. 1994. Issues of unity and objectivity. In (C. Peacocke, ed) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford University Press.
- Malpas, J. 1999. Constituting the mind: Kant, Davidson, and the unity of consciousness. *International Journal of Philosophical Studies* 7:1-30.
- Marcel, A. J. 1993. Slippage in the unity of consciousness. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.
- Marcel, A. J. 1994. What is relevant to the unity of consciousness? In (C. Peacocke, ed) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford University Press.
- Marks, C. 1980. *Commissurotomy, Consciousness, and Unity of Mind*. MIT Press.
- Marks, L. E. 1978. *The Unity of the Senses: Interrelations among the Modalities*. Academic Press.

- Maxwell, G. 1978. Unity of consciousness and mind-brain identity. In (J. Eccles, ed) *Mind and Brain*. Paragon House.
- McInerney, P. K. 1985. Person-stages and unity of consciousness. *American Philosophical Quarterly* 22:197-209.
- Metzinger, T. 1995. Faster than thought: Holism, homogeneity, and temporal coding. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.
- Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22:396-413. Reprinted in *Mortal Questions* (Cambridge University Press, 1979).
- Natsoulas, T. 1979. The unity of consciousness. *Behaviorism* 7:45-63.
- Natsoulas, T. 1984. Concerning the unity of consciousness: I. Varieties of conscious unity. *Imagination, Cognition and Personality* 3:281-303.
- Natsoulas, T. 1986. Concerning the unity of consciousness: II. William James on personal conscious unity. *Imagination, Cognition and Personality* 5:21-30.
- Oakley, D. A. & Eames, L. C. 1986. The plurality of consciousness. In (D. Oakley, ed) *Mind and Brain*. Methuen.
- O'Brien, G. & Opie, J. 1998. The disunity of consciousness. *Australasian Journal of Philosophy* 76:378-95.
- O'Brien, G. & Opie, J. 2000. Disunity defended: A reply to Bayne. *Australasian Journal Of Philosophy* 78:255-263.
- O'Shaughnessy, B. 1994. The diversity and unity of action and perception. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.
- Rosenberg, G. H. 1998. The boundary problem for phenomenal individuals. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness 1996*. MIT Press.
- Rossman, N. I. 1991. *Consciousness: Separation and Integration*. SUNY Press.
- Schleichert, H. 1985. On the concept of unity of consciousness. *Synthese* 64:411-20.
- Shoemaker, S. 1996. Unity of consciousness and consciousness of unity. In *The First-Person Perspective and Other Essays*. Cambridge University Press.

Stevenson, L. 2000. Synthetic unities of experience. *Philosophy and Phenomenological Research* 60:281-306.

Tinnin, L. 1990. Mental unity, altered states of consciousness, and dissociation. *Dissociation: Progress in the Dissociative Disorders* 3:154-59.

von der Malsburg, C. 1997. The coherence definition of consciousness. In (M. Ito, Y. Miyashita, & E. T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.

Ward, A. 1980. Materialism and the unity of consciousness. *Analysis* 40:144-46.

Watkins, J. W. N. 1982. A basic difficulty in the mind-brain identity hypothesis. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Zohar, D. 1995. A quantum-mechanical model of consciousness and the emergence of 'I'. *Minds and Machines* 5:597-607.

1.6c Homogeneity of Consciousness (Sellers, etc)

Clark, A. 1989. The particulate instantiation of homogeneous pink. *Synthese* 80:277-304.

Explains homogeneity in terms of nontransitive matching among pixelized parts of vision. Experience of continuity, not continuous experience. Experiences may have subphenomenal parts (e.g. invisible pixels).

Cornman, J. W. 1970. Sellers, scientific realism, and sensa. *Review of Metaphysics* 23:417-51.

Delaney, C. F. 1971. Sellers' grain argument. *Australasian Journal of Philosophy* 50:14-16.

Friedman, I. S. 1989. Ultimate homogeneity: A dialogue. *Philosophy Research Archives* 14:425-53.

Gunderson, K. 1974. The texture of mentality. In (R. Bambrough, ed) *Wisdom: Twelve Essays*. Blackwell.

Lockwood, M. 1993. The grain problem. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

Lycan, W. G. 1987. Sellers' "grain" argument. In *Consciousness*.

Metzinger, T. 1995. Faster than thought: Holism, homogeneity, and temporal coding. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Richardson, R. C. & Muilenberg, G. 1982. Sellers and sense impressions. *Erkenntnis*.

Sellars, W. S. 1963. Philosophy and the scientific image of man. In *Science, Perception, and Reality*. Humanities Press/Ridgeview.

Sellars, W. S. 1971. Seeing, sense impressions, and sensa: A reply to Cornman. *Review of Metaphysics*.

1.6d Knowledge of Consciousness [see also 1.3a, 1.8a, 5.13, 6.2i]

Alston, W. P. 1971. Varieties of priveleged access. *American Philosophical Quarterly* 8:223-41.

Alston, W. P. 1983. What's wrong with immediate knowledge? *Synthese* 55:73-96.

Armstrong, D. M. 1963. Is introspective knowledge incorrigible? *Philosophical Review* 62:417-32.

Armstrong, D. M. 1976. Incorrigibility, materialism, and causation. *Philosophical Studies* 30:125-28.

Bayne, T. 2001. Chalmers on the justification of phenomenal judgments. *Philosophy and Phenomenological Research* 62:407-19.

Bradley, R. D. 1964. Avowals of immediate experience. *Mind* 73:186-203.

Chalmers, D. J. 2002. The content and epistemology of phenomenal belief. In (Q. Smith & A. Jokic, eds) *Aspects of Consciousness*. Oxford University Press.

Chandler, J. H. 1970. Incorrigibility and classification. *Australasian Journal of Philosophy* 48:101-6.

Conee, E. 1994. Phenomenal knowledge. *Australasian Journal of Philosophy*.

Mary knew all the facts about qualia beforehand, she just wasn't acquainted with them. Knowledge by acquaintance isn't factual knowledge.

Dretske, F. 1999. The mind's awareness of itself. *Philosophical Studies* 95:103-24.

Dunlop, C. E. M. 1977. Lehrer and Ellis on incorrigibility. *Australasian Journal of Philosophy* 55:201-5.

Echelbarger, C. G. 1981. An alleged legend. *Philosophical Studies* 39:227-46.

Ellis, B. 1976. Avowals are more corrigible than you think. *Australasian Journal of Philosophy* 55:201-5.

Francescotti, R. M. 1994. Qualitative beliefs, wide content, and wide behavior. *Nous* 28:396-404.

Qualitative beliefs can supervene on behavioral dispositions even if absent/ /inverted qualia are possible.
We just individuate belief contents and behavior widely, with wide content fixed to the qualia.

Hill, C. S. 1988. Introspective awareness of sensations. *Topoi* 7:11-24.

Imlay, R. A. 1969. Immediate awareness. *Dialogue* 8:228-42.

Jackson, F. 1973. Is there a good argument against the incorrigibility thesis? *Australasian Journal of Philosophy* 51:51-62.

Kirk, R. 1971. Armstrong's analogue of introspection. *Philosophical Quarterly* 21:158-62.

Kornblith, H. 1998. What is it like to be me? *Australasian Journal of Philosophy* 76:48-60.

Leeds, S. 1993. Qualia, awareness, Sellars. *Nous* 27:303-330.

A discussion of in what sense we are aware of qualia, and how we can have beliefs about them, with reference to Sellars. Ends up reducing qualia to phenomenal beliefs in a language of thought. A rich and subtle paper.

Levin, J. 2001. The myth of Jones and the return of subjectivity. *Mind and Language* 16:173-192.

MacDonald, C. 1999. Shoemaker on self-knowledge and inner sense. *Philosophy and Phenomenological Research* 59:711-38.

Mackie, J. L. 1963. Are there any incorrigible empirical statements? *Australasian Journal of Philosophy* 41:12-28.

Margolis, J. 1970. Indubitability, self-intimating states, and privileged access. *Journal of Philosophy* 67:918-31.

Martin, M. 1998. An eye directed outward. In (C. Wright, B. Smith, and C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Nakhtnikian, G. 1968. Incorrigibility. *Philosophical Quarterly* 18:207-15.

Nida-Rumelin, M. 1995. What Mary couldn't know: Belief about phenomenal states. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Reconstructing the Mary case as a Marianna case, and introducing a distinction between phenomenal and nonphenomenal beliefs, to support the knowledge argument. A very nice and sophisticated paper.

Nida-Rumelin, M. 1997. On belief about experiences: An epistemological distinction applied to the knowledge argument. *Philosophy and Phenomenological Research*.

Odegard, D. 1992. Inner states. *Personalist Forum* 8:265-73.

Pappas, G. 1974. Incorrigibility, knowledge, and justification. *Philosophical Studies* 25:219-25.

Pappas, G. 1976. Incorrigibility and central-state materialism. *Philosophical Studies* 29:445-56.

Parsons, K. P. 1970. Mistaking sensations. *Philosophical Review*.

Peacocke, C. 1998. Conscious attitudes, attention, and self-knowledge. In (C. Wright, B. Smith, and C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Pollock, J. L. 1970. Perceptual knowledge. *Philosophical Review* 80:287-319.

Robinson, W. S. 1975. The legend of the given. In (H. Castaneda, ed) *Action, Knowledge, and Reality*. Bobbs-Merrill.

Robinson, W. S. 1982. Causation, sensation, and knowledge. *Mind* 91:524-40.

Argues that we can have non-inferential knowledge of sensations even if they don't make a causal difference to our beliefs, so epiphenomenalism is OK. All theories have a similar problems.

Rosenberg, J. 2000. Perception vs. inner sense: A problem about direct awareness. *Philosophical Studies* 101:143-160.

Schick, T. W. 1992. The epistemic role of qualitative content. *Philosophy and Phenomenological Research* 52:383-93.

Contra Sellars, Rorty, and Churchland: knowledge of qualitative content is an important aspect of our understanding of mental concepts, although it is not everything.

Sellars, W. 1956. Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science* 1:253-329. Reprinted as *Empiricism and the Philosophy of Mind*. Harvard University Press, 1997.

Shoemaker, S. 1990. First-person access. *Philosophical Perspectives* 4:187-214.

We have a limited special authority about the contents of our mental states. This follows from the link between a state and beliefs about it in the functional definition of that kind of state.

Smart, J. J. C. 1962. Brain processes and incorrigibility. *Australasian Journal of Philosophy* 40:68-70.

Reply to Baier 1962: epistemological authority is compatible with materialism. Mental state reports are not completely incorrigible, though.

Solomon, R. C. 1975. Minimal incorrigibility. *Australasian Journal of Philosophy* 53:254-56.

Sprigge, T. L. S. 1981. Knowledge of subjectivity. *Theoria to Theory* 14:313-25.

Tibbetts, P. 1972. Feigl on raw feels, the brain, and knowledge claims: Some problems regarding theoretical concepts. *Dialectica* 26:247-66.

Tomberlin, J. E. 1975. A problem with incorrigibility. *Philosophia* 5:507-12.

von Eckardt, B. 1975. Some consequences of knowing everything (essential) there is to know about one's mental states. *Review of Metaphysics* 29:3-18.

Wallraff, C. F. 1953. On immediacy and the contemporary dogma of sense-certainty. *Journal of Philosophy*.

Warner, R. 1993. Incorrigibility. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

Warner, R. 1994. In defense of a dualism. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Warner, R. 1996. Facing ourselves: Incorrigibility and the mind-body problem. *Journal of Consciousness Studies* 3:217-30.

1.6e The Function of Consciousness [see also 1.4e, 6.4a]

Baars, B. 1988. The functions of consciousness. In *A Cognitive Theory of Consciousness*. Cambridge University Press.

Argues for nine major functions of consciousness: in defining inputs, adaptation, debugging, recruiting & control, prioritizing, decision-making, analogy-forming, self-monitoring, and self-maintenance.

Banks, W. P. 1996. How much work can a quale do? *Consciousness and Cognition* 5:368-80.

Bechtel, W. & Richardson, R. C. 1983. Consciousness and complexity: evolutionary perspectives on the mind-body problem. *Australasian Journal of Philosophy* 61:378-95.

Contra Popper, evolution doesn't provide an argument against physicalism or epiphenomenalism.

Speculation on what the function of consciousness might be, and how it might be realized: e.g. selecting information.

Block, N. 1995. On a confusion about the function of consciousness. *Behaviora and Brain Sciences* 18:227-47.

Separates phenomenal consciousness from access consciousness, and argues that cases like blindsight only suggest a function for access consciousness, not phenomenal consciousness. The latter remains a mystery. With commentaries.

Bringsjord, S. & Noel, R. 1998. Why did evolution engineer consciousness? In (G. Mulhauser, ed) *Evolving Consciousness*.

DeLancey, C. 1996. Emotion and the function of consciousness. *Journal of Consciousness Studies* 3:492-99.

Dretske, F. 1997. What good is consciousness? *Canadian Journal of Philosophy* 27:1-15.

Flanagan, O. & Polger, T. 1995. Zombies and the function of consciousness. *Journal of Consciousness Studies* 2:313-21.

Argues for the possibility of zombies (contra Moody), then notes that any function could be performed by an unconscious zombie, it seems, so there's no function of consciousness in sight.

Flanagan, O. & Polger, G. 1998. Consciousness, adaptation, and epiphenomenalism. In (G. Mulhauser, ed) *Evolving Consciousness*.

Gregory, R. L. 1996. What do qualia do? *Perception* 25:377-79.

Suggests that qualia serve to distinguish hypotheses about present from past.

Kraemer, E. R. 1984. Consciousness and the exclusivity of function. *Mind* 93:271-5.

Contra Mott 1982: Function needn't be exclusive, and brain processes and consciousness may share a function, due to their close relationship.

McGinn, C. 1981. A note on functionalism and function. *Philosophical Topics* 12:169-70.

Function always underdetermines intrinsic nature, so absent/inverted qualia cases aren't incompatible with consciousness having a function.

Mott, P. 1982. On the function of consciousness. *Mind* 91:423-9.

Consciousness doesn't have a function, as any function it might have is a function of brain processes.

Perlis, D. 1997. Consciousness as self-function. *Journal of Consciousness Studies* 4:509-25.

Place, U. T. 2000. The causal potency of qualia: Its nature and its source. *Brain and Mind* 1:183-192.

Popper, K. R. 1978. Natural selection and the emergence of mind. *Dialectica* 32:339-55.

Ramachandran, V. S. & Hirstein, W. 1998. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies* 4:429-57.

Shanon, B. 1998. What is the function of consciousness? *Journal of Consciousness Studies* 5:295-308.

Tye, M. 1996. The function of consciousness. *Nous* 30:287-305.

Argues that the function of consciousness is not obvious, but that once one accepts a representational view of consciousness, it becomes obvious.

van Gulick, R. 1989. What difference does consciousness make? *Philosophical Topics* 17:211-30.

Trying to counter absent qualia arguments by finding a role for consciousness e.g. in metacognition, or as a way to achieve semantic transparency. But consciousness doesn't seem necessary for these, so it's still a mystery.

van Gulick, R. 1994. Deficit studies and the function of phenomenal consciousness. In (G. Graham & G. L. Stephens, eds) *Philosophical Psychopathology*. MIT Press.

Velmans, M. 1992. Is human information-processing conscious? *Behavioral and Brain Sciences* 14:651-69.

Uses experimental evidence to argue that consciousness is functionally inessential: the tasks associated with consciousness can be performed without consciousness. Only focal-attentive processing is required.

1.7 Qualia

1.7a

General

Burgess, J. A. 1990. Phenomenal qualities and the nontransitivity of matching. *Australasian Journal of Philosophy*.

Clark, Andy. 2000. A case where access implies qualia? *Analysis* 60:30-37.

Clark, A. 1985. Qualia and the psychophysical explanation of color perception. *Synthese* 65:377-405.

One can give an information-theoretic explanation of color perception, which leaves nothing out. Rebuts various qualia objections, e.g. from the possibility of inversion. Qualia are codes for external properties.

Clark, A. 1992. *Sensory Qualities*. Clarendon.

Argues that psychology is in the business of explaining sensory qualities, and does a perfectly good job using discriminability as a basis. With detailed argument and many interesting examples.

Clark, A. 2000. *A Theory of Sentience*. Oxford University Press

Cunningham, B. 2001. Capturing qualia: Higher-order concepts and connectionism. *Philosophical Psychology* 14:29-41.

Fox, I. 1989. On the nature and cognitive function of phenomenal content -- Part one. *Philosophical Topics* 17:81-103.

Searching for a theory of qualia: rejects epiphenomenalism, separation of the form and quality of experience, and immediate perception of phenomenal objects. Experience consists in represented (inexistent) objects of thought.

Gilbert, P. 1992. Immediate experience. *Proceedings of the Aristotelian Society* 66:233-250.

Against an account of phenomenal content as given by inner discrimination. Argues that the character of experience consists in its reason-giving role.

Gustafson, D. 1998. Pain, qualia, and the explanatory gap. *Philosophical Psychology* 11:371-387.

Hubbard, T. L. 1996. The importance of a consideration of qualia to imagery and cognition. *Consciousness and Cognition* 3:327-58.

Jakab, Z. 2000. Ineffability of qualia: A straightforward naturalistic explanation. *Consciousness and Cognition* 9:329-351.

Kind, A. 2001. Qualia realism. *Philosophical Studies* 104:143-62.

Kitcher, P. S. 1979. Phenomenal qualities. *American Philosophical Quarterly* 16:123-9.

Qualia problems stem from assuming direct awareness of perceptual states. Instead, we should acknowledge only an ability to detect and label these states. Also argues for the possibility of unconscious

and illusory pains.

Leeds, S. 1993. Qualia, awareness, Sellars. *Nous* 27:303-330.

A discussion of in what sense we are aware of qualia, and how we can have beliefs about them, with reference to Sellars. Ends up reducing qualia to phenomenal beliefs in a language of thought. A rich and subtle paper.

Leon, M . 1988. Characterising the senses. *Mind and Language* 3:243-70.

Levine, J. 1995. Qualia: Intrinsic, relational, or what? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Lormand, E. 1995. Qualia! (Now showing at a theater near you.) *Philosophical Topics* 22:127-156.

Mandik, P. 1999. Qualia, space, and control. *Philosophical Psychology* 12:47-60.

Nelkin, N. 1987. How sensations get their names. *Philosophical Studies* 51:325-39.

Sensations are an inessential element of experiences. Experiences are typed by their cognitive component, and the naming of sensations is derivative on this. With examples and empirical evidence about pain, color, perception.

Nelkin, N. 1990. Categorizing the senses. *Mind and Language*.

Putnam, H. 1981. Mind and body. In *Reason, Truth and History*. Cambridge University Press.

Considers qualia, inverted and absent, and various other stuff. Wishy-washy.

Rey, G. 1993. Sensational sentences. In (M. Davies & G. Humphreys, eds) *Consciousness: Philosophical and Psychological Essays*. Blackwell.

Explicating sensory experience in terms of an appropriate computational relation to a sentence in the language of thought. Argues that this handles many features of qualia (privacy, ineffability, grainlessness, unity, etc).

Robinson, W. S. 1999. Qualia realism and neural activation patterns. *Journal of Consciousness Studies* 10:65-80.

Shepard, R. N. 1993. On the physical basis, linguistic representation, and conscious experience of colors. In (G. Harman, ed) *Conceptions of the Human Mind: Essays in Honor of George A. Miller*. Lawrence Erlbaum.

Shoemaker, S. 1975. Phenomenal similarity. *Critica* 7:3-37. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Where does similarity come from? From belief therein? Similarity of experience = experience of similarity. Also relation to projectibility.

Shoemaker, S. 1994. Phenomenal character. *Nous* 28:21-38.

Phenomenal character is bestowed by representation of certain relational properties, defined by relation to experience. With a discussion of possible candidates, and argument against other views such as projectivism.

Sleutels, J. 1998. Phenomenal consciousness: Epiphenomenalism, naturalism and perceptual plasticity. *Communication and Cognition* 31:21-55.

1.7b

Qualia and Materialism [see also 1.2, 1.3]

Clark, A. 1985. A physicalist theory of qualia. *Monist* 68:491-506.

A Goodman-like theory of qualia discrimination.

Cornman, J. W. 1971. *Materialism and Sensations*. Yale University Press.

Double, R. 1985. Phenomenal properties. *Philosophy and Phenomenological Research* 45:383-92.

A somewhat vague defense of materialism against objections from phenomenal properties. The only problems are epistemological.

Harding, G. 1991. Color and the mind-body problem. *Review of Metaphysics* 45:289-307.

On the unique nature of color expanses, which are laid bare to perception as they are in themselves. These are incompatible with functionalist accounts of mind, but might still be physical, on a broader conception thereof.

Holborow, L. C. 1973. Materialism and phenomenal qualities. *Aristotelian Society Supplement* 47:107-19.

Horgan, T. 1987. Supervenient qualia. *Philosophical Review* 96:491-520.

Arguing from the causal efficacy of qualia and the closedness of physical causation to the conclusion that qualia conceptually supervene on the physical. A very thorough paper.

Jolley, K. D. & Watkins, M. 1998. What is it like to be a phenomenologist? *Philosophical Quarterly* 48:204-9.

A reply to Raffman 1995. Maybe our experiences are no more fine-grained than our concepts. Even our experiences of unique hues may be coarse.

Lewis, D. 1995. Should a materialist believe in qualia? *Australasian Journal of Philosophy* 73:140-44.

Materialists can believe in qualia, qua occupier of the folk psychological role. But they cannot accept the Identification Thesis, that having qualia allows us to know exactly what they are.

Lycan, W. G. 1988. Phenomenal objects: A backhanded defense. *Philosophical Perspectives* 3:513-26.

Argues that qualia, if viewed as simple properties of phenomenal individuals, are problematic for materialism. Considers the case for phenomenal individuals, and argues that they are intentional inexistents.

Marras, A. 1993. Materialism, functionalism, and supervenient qualia. *Dialogue* 32:475-92.

Qualia aren't reducible to physical properties, but they are supervenient (and ontologically dependent) on microfunctional properties. With remarks on the knowledge argument, Kripke, absent qualia, epiphenomenalism, etc.

Mellor, D. H. 1973. Materialism and phenomenal qualities II. *Aristotelian Society Supplement* 47:107-19.

Raffman, D. 1995. On the persistence of phenomenology. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Argues that our inability to reidentify qualia is a problem for materialism. How are they represented? Empty demonstrative would be vacuous, predicate would be reidentified, so maybe a plain presentation? Very interesting.

Tye, M. 1986. The subjective qualities of experience. *Mind* 95:1-17.

Absent/inverted qualia aren't really imaginable. The Knowledge Argument fails, as discovering new experiences doesn't imply learning new facts, but only coming to know old facts in a new way.

1.7c Eliminativism about Qualia [see also 1.4d]

Dennett, D. C. 1978. Why you can't make a computer that feels pain. *Synthese* 38. Reprinted in *Brainstorms* (MIT Press, 1978).

The concept of pain is incoherent, as it's asked to do too many things at once. With a discussion of drugs, flowcharts, reportability, etc.

Dennett, D. C. 1981. Wondering where the yellow went. *Monist* 64:102-8.

A response to Sellars. All there is to seeing occurrent yellow is the judgment that one is seeing occurrent yellow.

Dennett, D. C. 1988. Quining qualia. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Argues against the existence of ineffable, intrinsic, private, directly accessible properties. With lots of meaty-thought experiments, and arguments that there is no fact of the matter about inversion cases.

Dennett, D. C. 1991. Lovely and suspect qualities. In (E. Villanueva, ed) *Consciousness*. Ridgeview.

Everett, A. 1996. Qualia and vagueness. *Synthese* 106:205-226.

There are no qualia: qualia would have to be vague (for Sorites reasons), but there can be no vague properties in nature. The usual Sorites defenses don't work here, as there's no appearance/reality distinction for qualia.

Jacoby, H. 1985. Eliminativism, meaning, and qualitative states. *Philosophical Studies* 47:257-70.

Arguing against eliminativism for qualia. Even if nothing satisfies all the common-sense properties of qualia, reference of qualia terms is still fixed under a Putnam-style theory of meaning. Argues for scientific functionalism.

Levin, M. 1981. Phenomenal properties. *Philosophy and Phenomenological Research* 42:42-58.

There are no irreducible phenomenal properties. Materialism can handle our direct awareness of inner states by the right sort of causal connection. Gives a materialism account of discrimination and learning mental concepts.

Levine, J. 1994. Out of the closet: A qualophile confronts qualophobia. *Philosophical Topics* 22:107-126.

On bold vs. modest qualophilia, and against various qualophobic strategies. With remarks on scientific objectivity, qualia as an explanandum, and on how our knowledge of qualia is consistent with the conceivability of zombies.

Ross, D. 1993. Quining qualia Quine's way. *Dialogue* 32:439-59.

Seager, W. E. 1993. The elimination of experience. *Philosophy and Phenomenological Research* 53:345-65.

Dennett's 1988 argument against ineffability, etc., doesn't nearly make the case against qualia, and largely relies on verificationist assumptions.

Wright, E. W. 1989. Querying "Quining Qualia". *Acta Analytica* 4:9-32.

1.7d The Inverted Spectrum

Block, N. 1990. Inverted earth. *Philosophical Perspectives* 4:53-79.

Uses Inverted Earth case, colors and lenses inverted, to argue vs Harman that qualitative states aren't intentional states. Also, less convincingly, to argue that qualitative states aren't functional states.

Campbell, N. 2000. Physicalism, qualia inversion, and affective states. *Synthese* 124:239-256.

Casati, R. 1990. What is wrong in inverting spectra? *Teoria* 10:183-6.

Churchland, P. M. & Churchland, P. S. 1981. Functionalism, qualia and intentionality. *Philosophical Topics* 12:121-32. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Functional role counts more than qualitative content in determining what e.g. "redness" is.

Clark, A. 1985. Spectrum inversion and the color solid. *Southern Journal of Philosophy* 23:431-43.

Argues that there could be inverted spectra even without a symmetrical color space. Qualia must be distinguished from their place in color space.

Cole, D. J. 1990. Functionalism and inverted spectra. *Synthese* 82:207-22.

Acquired spectrum inversions do not refute functionalism, if qualia revert after behavioral adaptation (as they do with inverting lenses).

Dennett, D. C. 1994. Instead of qualia. In (A. Revonsuo & M. Kamppinen, eds) *Consciousness in Philosophy and Cognitive Neuroscience*. Lawrence Erlbaum.

Describes some "inverted spectrum" scenario in computer registers, and argues that in the absence of a "central clearing house", the inversion of qualia is indeterminate. There's no reason to believe in non-dispositional qualia.

Gert, B. 1965. Imagination and verifiability. *Philosophical Studies* 16:44-47.

Inverted spectra with constant behavior is a meaningful hypothesis even under verificationism. Switching nerve endings, tinting contact lenses, etc.

Hardin, C. L. 1987. Qualia and materialism: Closing the explanatory gap. *Philosophy and Phenomenological Research* 48:281-98.

On the physiological bases of phenomenal states, particularly color. Inverted spectrum isn't really coherent, as coolness/warmth would have to be inverted too. So the contingency of qualia is diminished.

Hardin, C. L. 1988. *Color for Philosophers*. Hackett.

Distinguishes various functionally distinct inverted spectrum cases.

Hardin, C. L. 1991. Reply to Levine. *Philosophical Psychology* 4:41-50.

Reply to Levine 1991. "Green residue" and "red residue" may be identical. Physiology might put more constraints on qualia, eventually ruling out all other possibilities. But there may still be absent/alien qualia problems.

Hardin, C. L. 1997. Reinverting the spectrum. In (A. Byrne & D. R. Hilbert, eds) *Readings on Color, Volume 1: The Philosophy of Color*. MIT Press.

Harrison, B. 1967. On describing colors. *Inquiry* 10:38-52.

Harrison, B. 1973. *Form and Content*. Blackwell.

The inverted spectrum is impossible, due to asymmetries in color space.

Harvey, J. 1979. Systematic transposition of colours. *Australasian Journal of Philosophy* 57:211-19.

The inverted spectrum can be detected, if a single person experiences both.

Hatfield, G. 1992. Color perception and neural encoding: Does metameric matching entail a loss of information? *Philosophy of Science Association* 1992, 1:492-504.

Johnsen, B. C. 1986. The inverted spectrum. *Australasian Journal of Philosophy* 64:471-6.

Against Shoemaker: physical realizations do not give empirical conditions for qualia inversion. Nice.

Johnsen, B. C. 1993. The intelligibility of spectrum inversion. *Canadian Journal of Philosophy* 23:631-6.

Kirk, R. 1982. Goodbye to transposed qualia. *Proceedings of the Aristotelian Society* 82:33-44.

The possibility of an inverted spectrum w.r.t. dispositions implies the falsity of physicalism. But this rests on an implausible "slide-viewer" model of seeing, and is incoherent otherwise.

Levine, J. 1988. Absent and inverted qualia revisited. *Mind and Language* 3:271-87.

Inverted qualia, with respect to a functional account, are no more plausible than absent qualia (by analysis of thought experiments). Both lead to first-person skepticism about qualia.

Levine, J. 1991. Cool red. *Philosophical Psychology* 4:27-40.

Contra Hardin 1988: there's a "green residue" after coolness is subtracted, so inverted spectrum could still be possible. In any case, the impossibility of IS doesn't affect the explanatory gap for qualia, which is epistemic.

Lycan, W. G. 1973. Inverted spectrum. *Ratio* 15:315-9.

Inverted spectrum holding behavior constant is at least a coherent idea. Hook up brain in different ways, etc.

Lycan, W. G. 1993. Functionalism and recent spectrum inversions. Manuscript.

Argues that qualia are intentional properties, and that inverted spectra, though conceivable, are metaphysically impossible, due to considerations about society and normality. Argues against Block's "inverted earth".

Nida-Rumelin, M. 1996. Pseudonormal vision: An actual case of qualia inversion? *Philosophical Studies* 82:145-57.

A fascinating note on the possibility of people with doubled colorblindness genes, thus inverting color processing; such people may actually exist.

Palmer, S. 1999. Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences*.

Rey, G. 1992. Sensational sentences reversed. *Philosophical Studies* 68:289-319.

Argues for a computational/sentential theory under which qualia are fixed by functional organization. Argues against Block's 1990 inversion: qualia might slowly change back as associations fade. Memory isn't 100% reliable.

Shoemaker, S. 1975. Phenomenal similarity. *Critica* 7:3-37. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Maybe IS is ongoing, with memory changes. What is the logic of "appears"?

Shoemaker, S. 1982. The inverted spectrum. *Journal of Philosophy* 79:357-381. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

All about the coherence and otherwise thereof. Uses switch in state for IS IS wrt behavior. Also claims that IS wrt function is possible as qualia are fixed by realizing state, not functional state. Bad assumption.

Shoemaker, S. 1996. Intersubjective/intrasubjective. In *The First-Person Perspective and Other Essays*. Cambridge University Press.

Taylor, D. 1966. The incommunicability of content. *Mind* 75:527-41.

Inverted spectra thought-experiments show that experiential content is incommunicable. Accounts for the fact that attempts to describe such cases lead to contradiction (I'm seeing green & not seeing green).

Tye, M. 1993. Qualia, content, and the inverted spectrum. *Nous*.

Argues that qualia are intentional properties, along the lines of "looks F to P". Handles inverted earth and related cases by taking the narrow intentional content. With remarks on the semantics of color terms.

1.7e

Absent Qualia (Block, etc) [see also 1.3b]

Block, N. & Fodor, J. A. 1972. What psychological states are not. *Philosophical Review* 81:159-81.

As a criticism of functionalism. raises the possibility that realizations of any given functional account of mental states may lack qualia.

Block, N. 1980. Troubles with functionalism. In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol 1. Harvard University Press.

All kinds of absent qualia cases: homunculi-headed robots, the population of China, and so on. There is a prima facie doubt that such cases lack qualia, so there is a prima facie case against functionalism.

Bogen, J. 1981. Agony in the schools. *Canadian Journal of Philosophy* 11:1-21.

It's OK for bizarre realizations to lack pain, as functionalism requires teleology as well as organization. With remarks on the relation between pain and "introspectible noxiousness".

Carleton, L. 1983. The population of China as one mind. *Philosophy Research Archives* 9:665-74.

Taking the personal stance, we should regard the Chinese nation as having qualia. A lack of qualia would make a functional difference.

Churchland, P. M. & Churchland, P. S. 1981. Functionalism, qualia and intentionality. *Philosophical*

Topics 12:121-32. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Absent qualia are impossible. Also, qualia aren't essential to mental state, functional role is.

Cuda, T. 1985. Against neural chauvinism. *Philosophical Studies* 48:111-27.

Replace neurons one by one with homunculi: what happens? Beliefs don't change, does consciousness fade? Very nice.

Elugardo, R. 1983. Functionalism, homunculi-heads and absent qualia. *Dialogue* 21:47-56.

If absent qualia are possible, then either qualia are inexplicable or species chauvinism is true. Homunculi-heads could make similar arguments about us.

Elugardo, R. 1983. Functionalism and the absent qualia argument. *Canadian Journal of Philosophy* 13:161-80.

Hardcastle, V. G. 1996. Functionalism's response to the problem of absent qualia. *Journal of Consciousness Studies* 3:357-73.

Jacoby, H. 1990. Empirical functionalism and conceivability arguments. *Philosophical Psychology* 2:271-82.

Conceivability arguments are only a problem for empirical functionalism insofar as they are a problem for materialism in general. Very true.

Juhl, C. F. 1998. Conscious experience and the nontriviality principle. *Philosophical Studies* 91:91-101.

Levin, J. 1985. Functionalism and the argument from conceivability. *Canadian Journal of Philosophy Supplement* 11:85-104.

Argues that metaphysical conclusions can be drawn from conceivability arguments, but that absent qualia cases have not been clearly and distinctly conceived. The functionalist is better off than the identity theorist here.

Levine, J. 1988. Absent and inverted qualia revisited. *Mind and Language* 3:271-87.

IQ are no more plausible than AQ (by analysis of thought experiments and skepticism). So there's no reason to choose physicalist-functionalism over pure functionalism, as Shoemaker does. Nice.

Sayan, E. 1988. A closer look at the Chinese Nation argument. *Philosophy Research Archives* 13:129-36.

The Chinese Nation would require less people than Churchland & Churchland 1981 suggest, as we'd only

need to handle a subset of all possible inputs.

Tye, M. 1993. Blindsight, the absent qualia hypothesis, and the mystery of consciousness. In (C. Hookway, ed) *Philosophy and the Cognitive Sciences*. Cambridge University Press.

Gives a thorough neurophysiological analysis of blindsight and related pathologies, and argues that these cannot be used to support the possibility of absent qualia. With remarks on the mystery of consciousness.

1.7f Introspection and Absent Qualia (Shoemaker) [see also 1.6d]

Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:291-315. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Absent qualia possible => qualia make no causal difference => no knowledge of qualia, therefore absent qualia are impossible. If qualia are introspectively accessible, they must be functional. An important argument.

Shoemaker, S. 1981. Absent qualia are impossible -- A reply to Block. *Philosophical Review* 90:581-99. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Reply to Block 1980. Distinguishes two AQ theses, and argues that if AQ are possible, then the problem for functionalism isn't due solely to qualia.

Averill, E. W. 1990. Functionalism, the absent qualia objection, and eliminativism. *Southern Journal of Philosophy* 28:449-67.

Defending Shoemaker's argument against Conee: immediate awareness and qualitative beliefs are the same. But maybe people *can't* tell whether they're having genuine or ersatz pain. Eliminativism is the best option.

Block, N. 1980. Are absent qualia impossible? *Philosophical Review* 89:257-74.

Reply to Shoemaker 1975. The possibility of absent qualia is compatible with a functional role for qualia, as qualia can make a causal difference that is independent of a given functional account.

Conee, E. 1985. The possibility of absent qualia. *Philosophical Review* 94:345-66.

Contra Shoemaker: qualia cause qualitative beliefs, which are affected by the absence of qualia, so we know about qualia even if AQ are possible.

Davis, L. 1982. Functionalism and absent qualia. *Philosophical Studies* 41:231-49.

Elucidating Shoemaker's argument: if absent qualia are possible, then the difference between real and ersatz pain makes no difference to belief, so qualia aren't introspectively accessible. A nice analysis.

Doore, G. 1981. Functionalism and absent qualia. *Australasian Journal of Philosophy* 59:387-402.

Qualia and qualitative beliefs are the same, so Shoemaker's argument fails. A numbness/pain inversion argument shows that pain isn't a functional state; it yields an introspectible difference without a functional difference.

Francescotti, R. M. 1994. Qualitative beliefs, wide content, and wide behavior. *Nous* 28:396-404.

Qualitative beliefs can supervene on behavioral dispositions even if absent/ /inverted qualia are possible. We just individuate belief contents and behavior widely, with wide content fixed to the qualia.

Hill, C. S. 1991. Introspection and the skeptic. In *Sensations: A Defense of Type Materialism*. Cambridge University Press.

Argues that the possibility of absent qualia is compatible with introspective knowledge. The fact that we have evidence of qualia isn't altered by the fact that we'd still think we had that evidence if we didn't have qualia.

White, N. 1985. Professor Shoemaker and the so-called 'qualia' of experience. *Philosophical Studies* 47:369-383.

Shoemaker's account leaves out experienced relations, such as experienced similarity. Experienced similarity is not the same as similarity between experiences. Being experienced is not an experienced feature.

1.7g Functionalism and Qualia, General

Brown, M. 1983. Functionalism and sensations. *Auslegung* 10:218-28.

Various comments on functionalism's troubles with qualia, including absent and inverted qualia. Analogies with biology and information theory.

Chalmers, D. J. 1995. Absent qualia, fading qualia, dancing qualia. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh. Press.

Argues that absent qualia and inverted qualia are empirically impossible (though logically possible), using neural-replacement thought-experiments. So functional organization fully determines conscious experience.

Cole, D. J. 1990. Functionalism and inverted spectra. *Synthese* 82:207-22.

Acquired spectrum inversion doesn't refute functionalism, if qualia revert after behavioral adaptation. With empirical evidence.

Dumpleton, S. 1988. Sensation and function. *Australasian Journal of Philosophy* 66:376-89.

Eshelman, L. J. 1977. Functionalism, sensations, and materialism. *Canadian Journal of Philosophy* 7:255-74.

Graham, G. & Stephens, G. 1985. Are qualia a pain in the neck for functionalists? *American Philosophical Quarterly* 22:73-80.

Pain-qualia are in the body, not the mind, and so aren't part of psychology.

Graham, G. & Stephens, G. 1987. Minding your P's and Q's: Pain and sensible qualities. *Nous* 21:395-405.

Greenberg, W. 1998. On Chalmers' "principle of organizational invariance" and his "dancing qualia" and "fading qualia" thought experiments. *Journal of Consciousness Studies* 5:53-58.

Hill, C. S. 1991. The failings of functionalism. In *Sensations: A Defense of Type Materialism*. Cambridge University Press.

Gives a number of arguments against both analytic functionalism and psychofunctionalism: arguments from absent qualia, absent functional role, epistemology, semantics, and heterogeneity of functional roles.

Horgan, T. 1984. Functionalism, qualia, and the inverted spectrum. *Philosophy and Phenomenological Research*, 44:453-69.

Argues that non-phenomenal mental events are functional, while qualia are low-level physiological.

Jarrett, G. 1996. Analyzing mental demonstratives. *Philosophical Studies* 84:49-62.

Lycan, W. G. 1981. Form, function and feel. *Journal of Philosophy* 78:24-50.

Accuses Block of a perspective error. Functionalism can handle a lot, if it's multi-levelled.

Lycan, W. G. 1987. Homunctionalism and qualia. In *Consciousness*. MIT Press.

Various stuff, mostly against absent qualia arguments.

Moor, J. H. 1988. Testing robots for qualia. In (H. Otto & J. Tuedio, eds) *Perspectives on Mind*. Kluwer.

Behavioral evidence for qualia is always indirect. And you can't check by replacing own neurons by chips, as you'll still believe you have qualia if you're functionally identical. Posit robot qualia as explanatory construct?

Nemirow, L. 1979. Functionalism and the subjective quality of experience. Dissertation, Stanford University.

Rey, G. 1994. Wittgenstein, computationalism, and qualia. In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*. Holder-Pichler-Tempsky.

Computational functionalism about qualia is compatible with Wittgenstein's views. It makes sense of the points about "dividing through" my private objects, for example. With remarks on spectrum inversions.

Seager, W. E. 1983. Functionalism, qualia and causation. *Mind* 92:174-88.

Functionalism can't explain the causal role of qualia by identifying them with functional states (circularity) or physical realizations (chauvinism). Which leaves property dualism, epiphenomenalism, or eliminativism for qualia.

Shoemaker, S. 1994. The first-person perspective. *Proceedings and Addresses of the American Philosophical Association* 68:7-22.

Against drawing strong conclusions from first-person imaginings. Considers Searle's silicon-replacement scenario: we might infer that perception isn't veridical, that there's another mind about, or even another body.

van Gulick, R. 1988. Qualia, functional equivalence and computation. In (H. Otto & J. Tuedio, eds) *Perspectives on Mind*. Kluwer.

Commentary on Moor 1988. Systems that differ in qualitative properties will likely differ in functional organization.

van Heuveln, B. , Dietrich, E. & Oshima, M. 1998. Let's dance! The equivocation in Chalmers' dancing qualia argument. *Minds and Machines*.

White, S. 1986. Curse of the qualia. *Synthese* 68:333-68.

Criticism of "physicalist-functionalism", where functional organization doesn't completely determine qualia (e.g. Shoemaker/Block). The only tenable options are pure functionalism or transcendental dualism. Nice.

White, S. 1989. Transcendentalism and its discontents. *Philosophical Topics* 17:231-61.

Taking transcendental dualism seriously. Privileged access provides strong arguments against objective theories, but it turns out that transcendentalism can't explain it any better, so maybe embrace objective theories after all.

Wright, E. 1995. More qualia trouble for functionalism: The Smythies TV-hood analogy. *Synthese* 97:365-82.

Zuboff, A. 1994. What is a mind? *Midwest Studies in Philosophy* 19:183-205.

Replacing a brain chunk while preserving causal role must preserve experience;

From chalmers Mon Jan 18 21:31:06 1999

Date: Mon, 18 Jan 1999 21:30:55 -0800 (PST)

From: David Chalmers <chalmers>

To: aburnette@u.arizona.edu, agillies@u.arizona.edu, akolers@u.arizona.edu, atlane@u.arizona.edu, bradt@u.arizona.edu, chalmers@ling.ucsc.edu, cowley@u.arizona.edu, erikh@u.arizona.edu, erikl@u.arizona.edu, jtismael@u.arizona.edu, kcreath@primenet.com, patrickr@u.arizona.edu, rachaelp@u.arizona.edu, sch@u.arizona.edu, shaughan@ns.arizona.edu, tbayne@u.arizona.edu, tolliver@u.arizona.edu

Subject: Philosophy 596B: Mind and Modality

Status: RO

To participants in Philosophy 596B (Mind and Modality):

As you probably know, I won't be in Tucson for the start of this course. Due to a bureaucratic mix-up, I wasn't granted a US visa in time for the start of term, and I'm currently in Australia (at ANU in Canberra) waiting for the visa to come through. I'm told that that should happen before February 15, and could happen as soon as the next few days, but there is no way to know. I hope that we won't miss more than two or three meetings of the seminar; in the worst case, we'll miss perhaps five.

Until I'm present in Tucson, the course will be conducted as an online seminar. The first meeting was scheduled for Tuesday Jan 19 (tomorrow Arizona time), so this can be considered a kick-off announcement. My plan is to make the first couple of weeks reading-intensive, with the opportunity for online questions and discussion. I'm currently trying to set up a listserv mailing list for the course and a web site; I'll let you know as soon as those things are operational. If it's possible, I'll try to arrange extra meetings once I make it to Tucson to get everything up to speed.

In the next message, I'll send a syllabus for the course, giving topics and readings and an approximate schedule. Note that the first part of the schedule will slightly be revised to accommodate the electronic format, to a degree that depends on the date of my return. In a third message (coming shortly) I'll detail initial topics, readings, and methodology.

Here's a list of participants in this seminar (the recipients of this mailing). If you know of someone else who should be on this list, please let me know. And if you shouldn't be on this list, let me know.

Enrolled students

Angela Burnette	aburnette@u.arizona.edu
Josh Cowley	cowley@u.arizona.edu
Anthony Lane	atlane@u.arizona.edu
Erik Larson	erikl@u.arizona.edu
Rachael Parkinson	rachaelp@u.arizona.edu
Brad Thompson	bradt@u.arizona.edu
Erik Herman	erikh@u.arizona.edu
Katherine Creath	kcreath@primenet.com

Auditors/sitting-in/interested:

Tim Bayne	tbayne@u.arizona.edu
Anthony Gillies	agillies@u.arizona.edu
Scott Hendricks	sch@U.Arizona.EDU
Avery Kolers	akolers@u.arizona.edu

Patrick Rysiew	patrickr@u.arizona.edu
Jenann Ismael	jtismael@u.arizona.edu
Shaughan Lavine	shaughan@ns.arizona.edu
Joseph Tolliver	tolliver@u.arizona.edu

Let me know if you have any questions. Note that the best e-mail address to use for me is my UCSC address, chalmers@ling.ucsc.edu.

--Dave Chalmers.

From chalmers Tue Jan 19 01:12:36 1999
Date: Tue, 19 Jan 1999 01:12:27 -0800 (PST)
From: David Chalmers <chalmers>
To: aburnett@u.arizona.edu, agillies@u.arizona.edu, akolers@u.arizona.edu,
atlane@u.arizona.edu, bayne@u.arizona.edu, bradt@u.arizona.edu,
chalmers@ling.ucsc.edu, cowley@u.arizona.edu, erikh@u.arizona.edu,
erikl@u.arizona.edu, jtismael@u.arizona.edu, kcreath@primenet.com,
laj@u.arizona.edu, patrickr@u.arizona.edu, rachaelp@u.arizona.edu,
sch@u.arizona.edu, shaughan@ns.arizona.edu, tolliver@u.arizona.edu
Subject: Initial readings and discussion
Status: RO

READING MATERIAL:

The reading material for the first bit of the course is:

The Conscious Mind, Chapters 2-4.
Jackson, Epiphenomenal Qualia [web].
Naming and Necessity.

You can also read Chapter 1 of TCM as background. But the course will very much be concentrating on the modal issues, so you might not want to become too distracted by general issues about consciousness.

Here's a suggested order of reading, corresponding roughly to the order in the syllabus. Because of the reading orientation, things will proceed a little differently from the way they would in person; for example, I'll leave some of the material intended for the first seminar until I get back. And you shouldn't think of the numbers below as corresponding to weeks, exactly. It might make sense to try to give most of this material a quick reading in the first week or so (assuming you haven't read it already, as some will have), before going back for a more detailed and careful reading to go along with a sequential discussion of these topics.

(1) Start with TCM sections 2.1-2.4 (pp. 32-55), for a general introduction to some issues about modality, ontology, and explanation. You might read the rest of 2.4 (pp. 55-71) and 2.5 lightly; we'll return to them later.

(2) Next read TCM Chapter 3 and section 4.1, as well as the Jackson piece, for an introduction to epistemic arguments against materialism.

(3) Next we come to a posteriori necessity. The relevant material here is _Naming and Necessity_, the rest of TCM section 2.4 (pp. 55-71, on the two-dimensional analysis of a posteriori necessity), and perhaps "The Components of Content" if you want more on the 2-D framework. The order of reading here depends to some extent on familiarity. If you're already familiar with the contents of N&N (as I assume many or most are), it would make sense to first read TCM 2.4, before going back to reread N&N in light of that analysis. If you haven't read N&N, you might want to read it first,

perhaps going back and forth with TCM 2.4 to gain a feeling for the 2-D analysis along the way.

Either way, the two-dimensional framework will be central to many topics in the course, so it's very important to gain a good understanding of it (I hope that before long you'll be living and breathing it). You'll want to go over that material very carefully to be sure that you understand it. "The Components of Content" has some further material (though applied to the semantics of thought rather than language) that may help you if you're not sure about things. After gaining a familiarity, try to translate all the central examples and arguments of N&N into the 2-D framework; that should prove to be a useful exercise in gaining an understanding. (I will probably make something like this an official assignment at some point.)

(4) Next (if we are still online) will be material on the application of the 2-D framework to the argument against materialism: the rest of TCM Chapter 4 (especially 4.2), and Jackson's "Finding the Mind in the Natural World". But I'll leave the details on that until later. If the seminar is in the flesh by this point, we'll have some material on subjunctive and indicative conditionals vis-a-vis the 2-D framework first.

ONLINE DISCUSSION:

We will start with a discussion of the material in (1) above, especially TCM 2.1. This is the closest thing easily available to an introduction to foundational issues concerning modality, ontology, and explanation. Try to get clear especially on the notions of logical and natural necessity, the various notions of supervenience, and the definition of materialism. The material on reductive explanation is secondary for our purposes, but you should still look at it. Post any questions and comments you might have to the list. Ideally, I'd like to see everyone (i.e. everyone officially enrolled in the course; participation by others is encouraged but optional) post some initial comments or questions in the next week (the sooner the better), so we can get an initial discussion going.

(N.B. I wouldn't now lay things out exactly as I do in those parts of TCM, but it should work as a first approximation. Hopefully a lot of things will be clarified and elaborated in discussion.)

I hope to have a listserv mailing list and/or web site for discussion up and running soon, but in the meantime, send all public messages to the list of recipients at the top of this message. N.B. My previous two messages had a couple of e-mail addresses wrong, so this is the one to use.

BACKGROUND:

It would be useful to get a sense of the class's familiarity with relevant issues. I'd appreciate it if everyone officially enrolled in the course could send me a note indicating their background in metaphysics, the philosophy of language, and the philosophy of mind: courses taken, relevant works read, and so on. It would be particularly useful to know whether and to what extent you're familiar with the ideas in and around (1) Naming and Necessity (2) The Conscious Mind (3) Lewis's On the Plurality of Worlds and (4) two-dimensional semantics. You might also mention your areas of main philosophical interest, your year, and any other relevant information, to help me get a sense of things. And you should feel free to take the opportunity to ask any questions that you might have. People who

aren't officially enrolled in the course should also feel free to send such a note.

(I was going to suggest sending such notes to the whole list, but don't want to clog people's mailboxes too badly, and some relevant information about background is probably best regarded as private. But people should feel free to send messages to the whole list to kick off discussion and to start building our little online community.)

--Dave.

From josh@ame2.math.arizona.edu Sun Jan 24 17:12:45 1999
From: Josh Cowley <josh@math.arizona.edu>
Subject: Re: Initial discussion
To: chalmers@ling.ucsc.edu (David Chalmers)
Date: Sun, 24 Jan 1999 18:11:56 -0700 (MST)
Cc: aburnett@u.arizona.edu, agillies@u.arizona.edu, akolers@u.arizona.edu, atlane@u.arizona.edu, bayne@u.arizona.edu, bradt@u.arizona.edu, chalmers@ling.ucsc.edu, cowley@u.arizona.edu, erikh@u.arizona.edu, erikl@u.arizona.edu, jtismael@u.arizona.edu, kcreath@primenet.com, laj@u.arizona.edu, patrickr@u.arizona.edu, rachael@u.arizona.edu, sch@u.arizona.edu, shaughan@ns.arizona.edu, tolliver@u.arizona.edu
Status: RO

Hi folks,

I guess I will get the ball rolling. I would like to note that I'm not really committed to anything I'm about to say.

I have some concerns about the use of "facts" and properties in the various definitions surrounding supervenience. My worry is pretty general, but I'll draw it out by focussing on the definition of a possitive fact or property. "A possitive fact in W is one that holds in everyw orld that contains W as a proper part; a positive property is one that if instantiated in a world W, is also instantiated by the corresponding individual in all worlds that contain W as a proper part." "(pg 40) In footnote 14 it is suggested that we ought to take the containment relation between worlds as primitive.

Here's my problem. Possible worlds are complicted and unnatural entities. Facts and properties are less complicated and more natural things. My inclination is to define possible worlds in terms of properties or facts. Something like, 'Two possible worlds, A and B, are identical just in case all the facts that hold in A hold in B and vice-versa. Two worlds differ in so far as the facts in A are not in B.'

Now there is nothing inherently wrong with defining possible worlds with facts and still saying that containment is a primitive notion. But it does seem to add some tension. I would be inclined to define containment by making reference to facts (or maybe properties) rather than using containment to define certain types of facts. What do people think?

Josh

From chalmers Wed Jan 27 04:24:41 1999
Date: Wed, 27 Jan 1999 04:24:28 -0800 (PST)
From: David Chalmers <chalmers>
To: aburnett@u.arizona.edu, agillies@u.arizona.edu, akolers@u.arizona.edu, atlane@u.arizona.edu, bayne@u.arizona.edu, bradt@u.arizona.edu,

chalmers@ling.ucsc.edu, cowley@u.arizona.edu, erikh@u.arizona.edu,
erikl@u.arizona.edu, jtismael@u.arizona.edu, laj@u.arizona.edu,
patrickr@u.arizona.edu, rachael@u.arizona.edu, sch@u.arizona.edu,
shaughan@ns.arizona.edu, tolliver@u.arizona.edu

Subject: Re: Initial discussion
Status: R0

Here's a reply to Josh's note. I'll hold off replying to Brad's note for a day or so to give other people a chance. I may not have said explicitly enough that everyone enrolled in the class should have posted some initial questions and comments by now. They don't have to be deep and profound; even simple clarifications or dumb questions are fine. But it's only through posting this sort of thing that we can get a good discussion going (hopefully on multiple fronts at once). We need to move on to the next material fairly soon. So I hope to see something from everyone within the next day or so.

Re Josh's point on the appeal to facts in the definition of containment. The worry is that if worlds are defined in terms of facts, then it may be troublesome to define positive facts in terms of worlds. Personally I'm neutral on whether possible worlds should be defined in terms of facts/properties or vice versa, but I'd like to think that even in the former case, the definition of positive fact is OK.

I guess the order of priority would be: (1) take the notion of a property (or a fact) as primitive; (2) build up the notion of a possible world out of these; (3) use the notion of a possible world so defined to characterize the class of positive facts. Of course one still needs to appeal an unreduced notion of containment relations between worlds; but it seems that we have clear intuitions about the application of this notion, and it doesn't seem especially problematic as unreduced notions go. Maybe Josh's idea is that it is less natural to apply an unreduced notion (containment) to an reduced ontology (worlds as facts) than to an unreduced ontology (worlds as basic), but it isn't clear to me that the former is problematic. Again, as long as we have clear judgments about its application.

I suppose the alternative would be to have a "fact-based" definition of a positive fact, or just to take the notion as primitive. I wouldn't wholly object to the latter, but personally I find the world-based definition (with containment as primitive) much more intuitive, and I think one ends up at least informally appealing to such a notion in explaining even a primitive notion of positive fact.

--Dave.

P.S. We are going to be particularly concerned with questions about the nature of and relationships between notions such as logical possibility and necessity, metaphysical possibility and necessity, a priori, a posteriori, natural possibility and necessity, etc. So any questions (including simple clarificatory questions) or comments there would be particularly welcome.

From bradt@u.arizona.edu Wed Jan 27 00:40:33 1999
From: "Brad Thompson" <bradt@u.arizona.edu>
To: "David Chalmers" <chalmers@ling.ucsc.edu>, <aburnett@u.arizona.edu>,
<agillies@u.arizona.edu>, <akolers@u.arizona.edu>,
<atlane@u.arizona.edu>, <bayne@u.arizona.edu>, <cowley@u.arizona.edu>,
<erikh@u.arizona.edu>, <erikl@u.arizona.edu>, <jtismael@u.arizona.edu>,
<kcreath@primenet.com>, <laj@u.arizona.edu>, <patrickr@u.arizona.edu>,
<rachael@u.arizona.edu>, <sch@u.arizona.edu>,

<shaughan@ns.arizona.edu>, <tolliver@u.arizona.edu>

Cc: <bradt@u.arizona.edu>

Subject: Re: Initial discussion

Date: Wed, 27 Jan 1999 01:34:18 -0700

X-MSMail-Priority: Normal

X-Priority: 3

Status: RO

Hi all!

Here's a question about the reading for last week. As I wrote this, I answered this question to my own satisfaction (the error I was making was obvious later)! But, in the interest of contributing to discussion and in the hopes that my own error will be illuminating to others, I will send the message anyways. No jeers or hate mail, please! :)

I find it difficult to conceive of 2 worlds that are identical with respect to their chemical facts but different with respect to their physical facts. In general, I'm worried that for all cases in which B-properties supervene on A-properties, it will turn out that A-properties supervene on B-properties. This would be odd, because it would entail that, for example, physical properties supervene on biological properties (since biological properties supervene on physical properties).

I think that the way that Dave specifies what it is for 2 worlds to be identical with respect to a set of properties is supposed to rule out this problem (see footnote 4). What we want is to be able to say that there are two possible worlds that are identical with respect to their biological properties but different with respect to their physical properties--this would rule out physical properties supervening on biological properties. So, let's imagine a biological duplicate of our world which, by stipulation, has a different physics. Wouldn't the structure and dynamics of that world's physics have to be the same as ours in order for it to have the same biological properties? But if so, it isn't clear that we've really imagined a world with a different physics.

Note: I don't think we can solve this problem simply by saying that there are possible worlds which are biologically identical and physically identical to our world except that there is extra physical stuff in that world. This would be analogous to arguing that biological facts don't supervene on the physical facts because of the possibility of a world with extra biological facts (like ectoplasmic angels).

* Here is where I went wrong. There *are* possible worlds with the same biological properties but with fewer or just different physical properties (as long as those absent or different physical properties belong to individuals which do not possess any biological properties). So, for example, there are possible worlds with the same biological properties as this one but in which my shirt is red rather than blue, or in which there simply is no Statue of Liberty.

Perhaps, despite this obvious blockade to the objection about supervenience above, my worry that a biologically identical world to this one would really have to have the same physical laws as this one is of independent interest. For example, if it makes sense at all to speak of a supervenience relation among types of properties rather than token instantiations of properties, the above consideration would suggest (maybe) that we would need to modify our definition of supervenience in a way that rules out physics supervening on biology. I'm not entirely sure this would be a consequence, since one would first need to figure out how we are to individuate or specify biological and physical property types.

From aburnett@U.Arizona.EDU Wed Jan 27 12:20:25 1999
Date: Wed, 27 Jan 1999 13:19:34 -0700 (MST)
From: Angela J Burnette <aburnett@U.Arizona.EDU>
To: Brad Thompson <bradt@U.Arizona.EDU>
cc: David Chalmers <chalmers@ling.ucsc.edu>, agillies@U.Arizona.EDU,
akolers@U.Arizona.EDU, atlane@U.Arizona.EDU, bayne@U.Arizona.EDU,
cowley@U.Arizona.EDU, erikh@U.Arizona.EDU, erikl@U.Arizona.EDU,
jtismael@U.Arizona.EDU, kcreath@primenet.com, laj@U.Arizona.EDU,
patrickr@U.Arizona.EDU, rachael@U.Arizona.EDU, sch@U.Arizona.EDU,
shaughan@ns.arizona.edu, tolliver@U.Arizona.EDU
Subject: Re: Initial discussion
Status: RO

All:

I have a question about supervenience...Brad suggested that we may have to modify our definition of supervenience to rule out the possibility of the physical supervening on the biological...On a related note, can someone please explain to me how the direction of a supervenience relation is built into the concept of supervenience? Or is this just a contingency that, once established, allows us to use the term when the appropriate relation is discovered? It just seems to me that there is nothing about the concept of supervenience that makes it move in one direction, and not both...so, my question is, what is it about supervenience that allows us to say that B supervenes on A without necessarily saying that A must also therefore supervene on B?

angela

From erikl@U.Arizona.EDU Wed Jan 27 12:31:58 1999
Date: Wed, 27 Jan 1999 13:30:16 -0700 (MST)
From: Erik J Larson <erikl@U.Arizona.EDU>
To: Brad Thompson <bradt@U.Arizona.EDU>
cc: David Chalmers <chalmers@ling.ucsc.edu>, aburnett@U.Arizona.EDU,
agillies@U.Arizona.EDU, akolers@U.Arizona.EDU, atlane@U.Arizona.EDU,
bayne@U.Arizona.EDU, cowley@U.Arizona.EDU, erikh@U.Arizona.EDU,
jtismael@U.Arizona.EDU, kcreath@primenet.com, laj@U.Arizona.EDU,
patrickr@U.Arizona.EDU, rachael@U.Arizona.EDU, sch@U.Arizona.EDU,
shaughan@ns.arizona.edu, tolliver@U.Arizona.EDU
Subject: Re: Initial discussion
Status: RO

Hi,
I'll try to reply to Brad and Angela on the question of the supervenience relation. The worry is that, if B-properties supervene on A-properties, then equally A-properties should supervene on B-properties. This seems, at best, odd--one doesn't want to say that physical facts supervene on biological facts or be committed to that view given that biological facts supervene on physical facts. I think the problem is that the notion of one set of facts "fixing" another set is left unanalyzed in the concept of a supervenience relation. It is not literally true that biological facts fix physical facts, because the physical facts have already fixed the biological facts, so there is no more work left to be done. but the notion of "fixed" seems a contingency left out of the strict definition of supervenience--at least insofar as we have encountered it in the readings thus far.

The other position is that, while physical facts MAY supervene on biological facts given the concept, such a statement is uninformative. But again, whether or not something is "informative" or not seems another

contingency that is left out of the concept of the supervenience relation.

What do you think?

Erik L.

From aburnett@U.Arizona.EDU Wed Jan 27 13:29:19 1999

Date: Wed, 27 Jan 1999 14:28:31 -0700 (MST)

From: Angela J Burnette <aburnett@U.Arizona.EDU>

To: Erik J Larson <erikl@U.Arizona.EDU>

cc: Brad Thompson <bradt@U.Arizona.EDU>,

David Chalmers <chalmers@ling.ucsc.edu>, agillies@U.Arizona.EDU,
akolers@U.Arizona.EDU, atlane@U.Arizona.EDU, bayne@U.Arizona.EDU,
cowley@U.Arizona.EDU, erikh@U.Arizona.EDU, jtismael@U.Arizona.EDU,
kcreath@primenet.com, laj@U.Arizona.EDU, patrickr@U.Arizona.EDU,
rachael@U.Arizona.EDU, sch@U.Arizona.EDU, shaughan@ns.arizona.edu,
tolliver@U.Arizona.EDU

Subject: Re: Initial discussion

Status: RO

all,

regarding Erik's take on supervenience, that sounds good but, given that the mentioned aspects are left out of the concept of supervenience then we are just using the word as a convenient label? I think that I was categorizing supervenience alongside words like entailment, for example. Establishing that an entailment holds between A and B allows you to do all sorts of things in virtue of the entailment, but if supervenience is just a label, then it won't be helpful in establishing new facts or conclusions about anything in the way that a concept like entailment does, right?

angela

From chalmers Wed Jan 27 19:27:33 1999

Date: Wed, 27 Jan 1999 19:27:23 -0800 (PST)

From: David Chalmers <chalmers>

To: chalmers

Subject: Supervenience and symmetry

Cc: aburnett@u.arizona.edu, agillies@u.arizona.edu, akolers@u.arizona.edu,
atlane@u.arizona.edu, bayne@u.arizona.edu, bradt@u.arizona.edu,
chalmers@ling.ucsc.edu, cowley@u.arizona.edu, erikh@u.arizona.edu,
erikl@u.arizona.edu, jtismael@u.arizona.edu, laj@u.arizona.edu,
patrickr@u.arizona.edu, rachael@u.arizona.edu, sch@u.arizona.edu,
shaughan@ns.arizona.edu, tolliver@u.arizona.edu

Status: RO

Some very interesting points re supervenience. Angela asks about whether directionality is built into the supervenience concept. I would say that it is built in, as the definition is asymmetrical. Take the first definition: B supervenes on A if all A-indiscernible worlds are B-indiscernible. One can see that the definition is asymmetrical by noting that if the A-indiscernible worlds are a proper subset of the B-indiscernible worlds, B will supervene on A but not vice versa. Something similar applies to the modified definitions given later on.

So there's at least conceptual room for B to supervene on A but not vice versa. And it's not too hard to find examples. Let A be exact height, and let B be the property of being over six feet tall. Then B supervenes on A but not vice versa.

Brad raises the concern that in the case where the A-properties are the physical properties (the usual case for our purposes), it may be hard to find asymmetries. I think one can find at least a few straightforwardly asymmetrical cases by taking e.g. B = the total mass of the universe, or B = the property of having mass, or even B = position. It wouldn't be too hard (I think) to construct a world that is physically different from ours, but has the same total mass, or has the same distribution of massive objects (but e.g. with different precise masses), or has the same distribution of positions (but maybe different particles at some positions).

The physics/chemistry and physics/biology cases are at least somewhat trickier, though. Of course things depend on just what we take to be the chemical and biological properties. But there are at least two natural places to look for asymmetry.

(i) Entities in our world that have physical properties but no biological (chemical) properties, and which make no contribution to biological (chemical) properties of other entities. Maybe isolated photons, for example. Then there's presumably a possible world a lot like ours but where that entity is absent. That world will be biologically but not physically identical to ours. Brad gives some cases like this.

(ii) Entities with biological properties and some "biologically irrelevant" physical properties. E.g. one *could* claim that one could change the position of one particle in my big toenail an iota without changing any of my biological properties. Of course that depends on just what we are counting as a biological property.

Anthony also gives the example of switching physical properties wholesale (e.g. matter to anti-matter) while preserving biological properties.

[Anthony also asks whether natural supervenience can be similarly asymmetrical. I think the answer is that it is a bit harder to find asymmetries, because we have fewer possible worlds to play with, but we should still be able to find them. E.g. the world without the photon, or with the moved particle in my toenail, are presumably naturally possible. And clearly for examples such as the height case, one has asymmetrical natural supervenience. Finally, someone like me might argue that the psychophysical case is an example of asymmetric natural supervenience, if e.g. a silicon duplicate of me might have the same phenomenal properties -- abstracting away from certain worries posed by panpsychism and the like.]

This being said, presumably there are certain liberal ways of construing biological properties that may lead to asymmetry. For example, if one includes things like "being exactly 2.3 light years from something that is alive", and similar properties, then even the photon case might come out symmetrical. I recall there was a paper in the *Journal of Philosophy* a few years back:

Miller, R.B. 1990. Supervenience is a two-way street. *Journal of Philosophy* 87:695-701.

that made something like this point to argue that most interesting cases of supervenience are symmetrical.

Personally I think it makes sense to go with a narrower notion of biological property. But this sort of example does at least make the

useful point that the mere fact that B supervenes on A doesn't show that B is ontologically dependent on A. In the case above, the physical properties might supervene on fine-grained and relational biological properties, presumably without ontologically depending on them. Some people react by insisting that we need a further clause of "ontological dependence" for the supervenience relation to be useful. I think Grimes made a point like this in

Grimes, T. 1988. The myth of supervenience. Pacific Philosophical Quarterly 69:152-60.

arguing that supervenience isn't at all useful in understanding ontological dependence. Some similarly argue that at least an asymmetry clause should be built into the definition.

Personally I think we should leave the definition of supervenience as it is, allow the conceptual possibility that A and B can supervene on each other in some cases, and consider the question of ontological dependence case by case. I think one can at least say that when (logical/metaphysical) supervenience fails, ontological dependence fails, so supervenience can be considered a necessary condition here. Just what further is required to give a sufficient condition is an interesting question. Any ideas are welcome! (One might presumably also want to deal with questions about e.g. mathematical facts supervening on anything (as they're necessarily true) without being ontological dependent on those things, etc.)

I think Erik L.'s point about fixing and informativeness is related to this point. I think Erik is pointing toward the sense of "fixes" in which A "fixes" B when B ontologically depends (and is determined by) A. So in the cases above, physics might supervene on fine-grained biology without being "fixed" by it in the ontological sense (though I think one could say that it is "fixed" at least in a modal sense). Again, this points to the fact that supervenience doesn't give a complete accounting of ontological dependence. Similarly, one can argue that it doesn't give a complete account of reductive explanation (I think Erik's point about "informativeness" is tied to this). So again people could think about what we need to add to supervenience to get ontological dependence and reductive explanation.

I think in most cases of supervenience on the physical, it's plausible that this goes hand in hand with ontological dependence on the physical, though there may be a tricky case or two we'll encounter later on. (E.g., what if a strong sort of panpsychism is true, and every physical particle is constituted by an inner phenomenal essence?) In any case the central application of supervenience will be in its guise as a necessary condition for ontological dependence; e.g. in the mind-body case where a failure of supervenience would imply the falsity of materialism. The idea that supervenience is a necessary condition for ontological dependence is relatively uncontroversial, I think, at least among those who buy into modal talk about these things.

All this ties into Angela's second point about supervenience as a convenient label. It's certainly the case that supervenience won't solve all our problems for us; it may not give an exhaustive analysis of ontological dependence or reductive explanation, for example. On the other hand, it may be able to do a good amount of work in clarifying those notions (e.g. at least as a necessary conditions, and perhaps as part of a sufficient condition); and it certainly has roles in other philosophical areas. I guess the moral is that a technical notion such as supervenience should be seen as our servant, not as our

master. (I think more or less the same goes for notions such as "entailment", actually.) It doesn't solve major philosophical problems for us on its own, but we can hope that it at least plays a useful role in clarifying them, and we can judge its application and usefulness on a case-by-case basis as we go.

That's enough for one note. All thoughts are welcome!

--Dave.

P.S. I've used Thony's idea of putting the mailing list in the Bcc label. The only trouble is that it means one can't reply directly to the list. That isn't a problem as long as one has kept a copy of the list around the place. If anyone needs a copy, let me know. Hopefully a listserv will be set up soon.

From chalmers Wed Jan 27 20:07:51 1999
Date: Wed, 27 Jan 1999 20:07:47 -0800 (PST)
From: David Chalmers <chalmers>
To: chalmers
Subject: Laws, angels, etc.
Status: RO

A few other issues that came up.

(1) Is there a strict boundary between logical and natural possibility, or a gray zone? (Erik H.)

Hmm. I suppose I'd say there's a gray zone for logical possibility at least of statement, because of vagueness in the meaning of words. E.g., it might be vague whether it is logically possible to have a heap with four grains of sand. But that's not a vagueness in the boundary between logical and natural possibility per se. A vagueness in that boundary would require something that is clearly logically possible, such that it's vague whether it is naturally possible. I suppose one could again find examples due to vagueness in language (say it were vague whether photons counted as "particles"; then it might be vague as to whether massless particles are naturally possible). But those aren't very interesting. Leaving such cases aside, a gray zone for natural possibility would presumably involve a gray zone in the laws of nature themselves. I suppose one could try to find such areas of grayness -- e.g. sometimes it's not clear how much variation in initial conditions should be counted as naturally possible, and how much should be seen as built into the laws of nature. But I think for most of our purposes it won't hurt to see the bounds of natural possibility as fairly strict.

(2) What if there are nonphysical angels (etc.) in our world? (Erik H.)

If that's the case, then materialism is certainly false in our world. Here the supervenience definition in 2.1 seems to give the right result. There's a world physically identical to ours but without the angels (presumably), so angels won't logically supervene on the physical by the account on p. 39 (and p. 40). And the angel facts will be positive facts about our world that don't supervene, so materialism will be false by the account on p. 41.

Of course, we don't need to make assumptions about whether or not there are angels in our world for the purposes of defining materialism, as long as our definition gives the right results either way. The particular difficulty is raised by the case in which (1)

there are no angels (etc.) in our actual world, but (2) there's a physically identical possible world (not actual) with angels. In this case, the intuition is that materialism should be true in our world, despite the existence of a physically identical world where things differ. Hence the modifications to the definition. In the case where the actual world has angels, the original definition was already OK.

Re angels traveling across worlds: I guess I'd say (with Lewis) that if a being can move from one world to another, then those worlds are causally connected, and so aren't really different worlds in the first place. Just one big world with two parts. Of course the same being might arguably be present in two worlds (e.g. a possible world where I'm a mathematician, not a philosopher), but that's not the same as traveling between worlds.

(3) Are laws of nature captured in the physical properties of the world, or are they additional? (Anthony.)

This comes down to the question of Humean vs non-Humean views of laws. Personally I lean toward a non-Humean view, holding that laws involve more than any kind of regularity. It seems to me that two physically identical worlds (at least identical in their distribution of physical properties) could have different laws (e.g., Tooley's example of worlds that differ in the laws governing the behavior of a pair of particles that never actually meet), and maybe that there could be a world physically identical to ours but with no real laws at all (the Hume world where it's a giant cosmic coincidence). If so, laws (and causation, etc.) involve something ontologically significant in the world over and above distribution of particular physical properties. I discuss this a little on p. 86 of TCM.

I think a commitment on this view isn't essential to most of the core issues we'll be discussing, though. I avoid the issue for the most part in the book by stipulating that physically identical worlds have the same physical laws (i.e., physical laws are built into the physical supervenience base). The issue is interesting, though, as another example of the way in which ontology and modality interact, and as another domain in which there are modal arguments with ontological consequences. We'll see later on that there are interesting parallels between the law case and the phenomenal case.

One complication is that a smallish minority of philosophers hold that laws of nature are metaphysically necessary and hold in all possible worlds. If that's so, then metaphysical and natural necessity arguably collapse (and so will metaphysical and natural supervenience, etc). On my preferred usage, such laws still won't be logically necessary, as it is still conceivable that they are false, there are conceivable scenarios with different laws, etc; so we'll still have a distinct notion of logical necessity and logical supervenience. (The necessity of laws will be an example of what I call in the book "strong metaphysical necessities".) But this is a tricky matter that we'll be discussing later on. (Again, any clarifying questions on logical vs. metaphysical necessity, etc, are welcome.)

(4) Thony's (interesting) question re Kripke and the 2-D framework. Maybe I'll save that one until we're focusing on those issues in a little while.

Looking forward to people jumping in and continuing the discussion...

--Dave.

P.S. I hope I really Bcc'd it this time.

From erikh@U.Arizona.EDU Wed Jan 27 08:13:47 1999
Date: Wed, 27 Jan 1999 09:12:59 -0700 (MST)
From: Erik A Herman <erikh@U.Arizona.EDU>
To: David Chalmers <chalmers@ling.ucsc.edu>
cc: aburnett@U.Arizona.EDU, agillies@U.Arizona.EDU, akolers@U.Arizona.EDU,
atlane@U.Arizona.EDU, bayne@U.Arizona.EDU, bradt@U.Arizona.EDU,
cowley@U.Arizona.EDU, erikl@U.Arizona.EDU, jtismael@U.Arizona.EDU,
laj@U.Arizona.EDU, patrickr@U.Arizona.EDU, rachaelp@U.Arizona.EDU,
sch@U.Arizona.EDU, shaughan@ns.arizona.edu, tolliver@U.Arizona.EDU
Subject: Re: Initial discussion
Status: RO

OK, here goes,

First is a question: by definition, is there a strict boundary between logical and natural possibility, or is there a gray zone?

Comment: The idea that there is a world physically identical to ours but has extra nonphysical stuff: Can we postulate anything about that world when we don't know about our own worlds non-physical stuff--if it is bound by logical entailments, etc? (if that makes any sense) Assuming our world has no angels is a strong assumption perhaps. And maybe these angels can travel across worlds? Is that allowed?

-Erik

From atlane@U.Arizona.EDU Wed Jan 27 14:08:55 1999
Date: Wed, 27 Jan 1999 15:08:10 -0700 (MST)
From: Anthony T Lane <atlane@U.Arizona.EDU>
To: Erik J Larson <erikl@U.Arizona.EDU>
cc: Brad Thompson <bradt@U.Arizona.EDU>,
David Chalmers <chalmers@ling.ucsc.edu>, aburnett@U.Arizona.EDU,
agillies@U.Arizona.EDU, akolers@U.Arizona.EDU, bayne@U.Arizona.EDU,
cowley@U.Arizona.EDU, erikh@U.Arizona.EDU, jtismael@U.Arizona.EDU,
kcreath@primenet.com, laj@U.Arizona.EDU, patrickr@U.Arizona.EDU,
rachaelp@U.Arizona.EDU, sch@U.Arizona.EDU, shaughan@ns.arizona.edu,
tolliver@U.Arizona.EDU
Subject: Re: Initial discussion
Status: RO

Erik points out that if B-properties supervene on A-properties, it seems that A-properties must also supervene on B-properties. It seems that this possibility depends on the type of supervenience being discussed. If we are talking about logical supervenience, this does not seem to be true. At the global level, given that the world has certain physical properties and natural laws that constitute its A-properties, there is no possible way in which it could have different biological properties. However, the reverse of this does not seem to be true-- one can imagine a world that is biologically identical to this one, but in which there are slightly different physical properties. For instance, one can imagine a world composed of anti-matter, but which has suitably different laws than we have in our own world such that it is biologically indistinguishable from our own world.

I am not sure whether this is also true in the case of natural supervenience. Is it possible that, given the natural laws that hold in our world, a world could be biologically identical to ours and yet be physically different? I would hazard a guess that it might be. We can imagine a world identical to our own biologically, except that organisms

are based on Silicon instead of Carbon. Of course, it might be impossible that such a world could be biologically identical to this one-- it would have to be a world in which there is a great abundance of Silicon and it must somehow behave chemically just like Carbon does in this world. Perhaps this is ruled out by the laws of nature.

A question about laws of nature: are they somehow captured in an exhaustive list of the physical properties of the world, or are they additional facts that are added on? I expect you (Dave) do not need to say much about them beyond the fact they may be different in logically possible worlds, but not in naturally possible worlds.

Anthony

From rachael@U.Arizona.EDU Thu Jan 28 09:29:31 1999
Date: Thu, 28 Jan 1999 10:28:46 -0700 (MST)
From: Rachael J Parkinson <rachael@U.Arizona.EDU>
To: David Chalmers <chalmers@ling.ucsc.edu>
cc: aburnett@U.Arizona.EDU, agillies@U.Arizona.EDU, akolers@U.Arizona.EDU, atlane@U.Arizona.EDU, bayne@U.Arizona.EDU, bradt@U.Arizona.EDU, cowley@U.Arizona.EDU, erikh@U.Arizona.EDU, erikl@U.Arizona.EDU, jtismael@U.Arizona.EDU, laj@U.Arizona.EDU, patrickr@U.Arizona.EDU, sch@U.Arizona.EDU, shaughan@ns.arizona.edu, tolliver@U.Arizona.EDU
Subject: Re: Supervenience and symmetry
Status: RO

David has suggested that global-logical supervenience is necessary but not sufficient for reductive explanation. It seems clear that supervenience is a necessary condition for reductive explanation- but it is not so clear that a good explanation of a phenomenon can be cashed out merely in terms of the lower-level facts. Searle implies this when he suggests that it is inadequate to explain a cocktail party or a football game merely in terms of the physical facts, though logical supervenience implies that this is possible. If a reductive explanation is to be illuminating, perhaps what is necessary beyond supervenience is what David suggested on page 51. He suggests that the project of reductive explanation requires both explication and explanation. Thus we can explicate a football game in terms of the rules, the players, the plays, etc. and explain it in terms of the lower-level physical facts, eliminating any mystery surrounding the high-level phenomena. My question: can some one expand on why local logical supervenience is too stringent a requirement for reductive explanation?(51) I'm not disputing the claim- I would just like to understand it better.
-Rachael

From erikl@U.Arizona.EDU Thu Jan 28 10:24:39 1999
Date: Thu, 28 Jan 1999 11:22:59 -0700 (MST)
From: Erik J Larson <erikl@U.Arizona.EDU>
To: Rachael J Parkinson <rachael@U.Arizona.EDU>
cc: David Chalmers <chalmers@ling.ucsc.edu>, aburnett@U.Arizona.EDU, agillies@U.Arizona.EDU, akolers@U.Arizona.EDU, atlane@U.Arizona.EDU, bayne@U.Arizona.EDU, bradt@U.Arizona.EDU, cowley@U.Arizona.EDU, erikh@U.Arizona.EDU, jtismael@U.Arizona.EDU, laj@U.Arizona.EDU, patrickr@U.Arizona.EDU, sch@U.Arizona.EDU, shaughan@ns.arizona.edu, tolliver@U.Arizona.EDU
Subject: Re: Supervenience and symmetry
Status: RO

Rachel, the key is "local" logical supervenience. Dave gave the example of two physically identical organisms (ie meeting the condition of local

supervenience) with non-supervenient high-level properties (their biological "fitness" say) because of differences in the match of the physical characteristics of the organisms to their environment. So if physical facts are not considered globally, certain facts about individuals may fail to supervene, even though they are entirely physical. I want to say that certain properties about individuals are context-dependent in a way that requires global fixing of facts to cache out the supervenience relation. Searle's quip about the football game is consistent with this.

Erik L.

On Thu, 28 Jan 1999, Rachael J Parkinson wrote:

> David has suggested that global-logical supervenience is necessary but not
> sufficient for reductive explanation. It seems clear that supervenience is
> a necessary condition for reductive explanation- but it is not so clear
> that a good explanation of a phenomenon can be cached out merely in terms
> of the lower-level facts. Searle implies this when he suggests
> that it is inadequate to explain a cocktail party or a football game
> merely in terms of the physical facts, though logical supervenience
> implies that this is possible. If a reductive explanation is to be
> illuminating, perhaps what is necessary beyond supervenience is what David
> suggested on page 51. He suggests that the project of reductive
> explanation requires both explication and explanation. Thus we can
> explicate a football game in terms of the rules, the players, the plays,
> etc. and explain it in terms of the lower-level physical facts,
> eliminating any mystery surrounding the high-level phenomena.
> My question: can some one expand on why local logical supervenience is too
> stringent a requirement for reductive explanation?(51)I'm not disputing
> the claim- I would just like to understand it better.
> -Rachael
>
>

From chalmers Thu Jan 28 17:54:08 1999
Date: Thu, 28 Jan 1999 17:54:05 -0800 (PST)
From: David Chalmers <chalmers>
To: chalmers
Subject: Re: Supervenience and symmetry
Status: RO

Rachel is right that logical supervenience alone doesn't capture everything that is relevant about explanation. I do think that it's captures what's going on with reductive explanation, especially when one thinks of this as a two-stage process of explication plus empirical satisfaction. But it's also true that reductive explanation is far from the be-all and end-all of explanation. Obviously a reductive explanation of a football game or a cocktail party may not provide the most illuminating explanation of what is going on there. It will be a "mystery-removing" explanation in that it will explain how football games and cocktail parties are possible, but even after the process of explication plus empirical satisfaction is complete, there will presumably be a major role for high-level explanations in helping us to understand the high-level phenomena in a comprehensible way.

There is still the point which came up last time, concerning cases (e.g. symmetrical cases) where B supervenes on A while not being intuitively ontologically dependent on A, or reductively explainable in terms of A. The supervenience of the physical upon sufficiently

fine-grained biological properties might be an example. Maybe to handle this sort of case one needs to build in some further clause about the autonomy of the domain of the supervenience base. It's not entirely obvious how to cash this out, though, and all ideas are welcome.

Re local supervenience, I think we often want to say that phenomena are reductively explainable even when they aren't locally supervenient. Fitness is an example, as Erik suggests. Perhaps my being a brother is another one. These things don't supervene locally, as they depend constitutively on environmental factors. But it still seems reasonable to say that we can give reductive explanations of them. In such a case, a reductive explanation will advert to lower-level properties of both me (locally) and of my environment.

--Dave.

From agillies@U.Arizona.EDU Mon Feb 1 11:04:11 1999
X-Sender: agillies@pop.u.arizona.edu
Date: Mon, 1 Feb 1999 11:49:16 -0700
To: Recipient.List.Suppresse:;
From: "Anthony S. Gillies" <agillies@U.Arizona.EDU>
Subject: discussion: worry about epistemic arguments
Status: RO

Hi all,

I have a sort of ill-defined, amorphous worry about epistemic arguments against materialism. But, be warned, the worry seems to come and go as it pleases, and I may not have it by the time I get to the end of this message.

In general epistemic arguments against materialism have the form:

1. Imagine someone (typically someone named 'Mary') who knows all there is to know about X (color vision, etc.)
2. But, imagine Mary has cognitive deficiency Y (color blind, raised in a black and white room, etc.)
3. One day she's released (or her color vision miraculously is restored, or ...) and sees *green* grass for the very first time.
4. She's learned something new; ergo materialism is false.

Here's my worry (I think). (Incidentally, it probably rests on another amorphous, ill-defined worry about conceivability.) How can we be sure we're getting our inferences about Mary's inferences right? The argument needs it to be that Mary learns nothing new, but as far as I can tell this requires us to track Mary's inferencing. But this is hard for us since *we don't know* all there is to know about X. And how much someone knows about X has an obvious impact on the conclusions one can draw, no?

Here's an example of what I have in mind. The Churchland's have a pretty cool parody of a thought experiment called (I think) the "Dark Room Argument". I'll change their example to make it an epistemic argument.

1. Imagine Mary knows all there is to know about light.
2. But Mary is in a completely dark room.
3. Suppose Mary is swinging a bar magnetic abover her head in the dark room. Nothing happens. Then someone turns on a light.
4. She's learned something new: light isn't electromagnetic wave propagation after all.

Now, since we know more about light and how it worls, the argument is unconvincing. But we can imagine that it may have been compelling to (say) folks in the middle ages.

Like I said, the worry comes and goes. But at the moment I am a bit worried about this sort of thing.

Thony

+++++

"Curious green ideas sleep furiously."

+++++

From josh@ame2.math.arizona.edu Mon Feb 1 12:23:06 1999
From: Josh Cowley <josh@math.arizona.edu>
Subject: epist. and analysis arguments
To: bradt@U.Arizona.EDU (Brad Thompson)
Date: Mon, 1 Feb 1999 13:22:21 -0700 (MST)
Cc: chalmers@ling.ucsc.edu, aburnett@U.Arizona.EDU, agillies@U.Arizona.EDU, akolers@U.Arizona.EDU, atlane@U.Arizona.EDU, bayne@U.Arizona.EDU,

cowley@U.Arizona.EDU, erikh@U.Arizona.EDU, erikl@U.Arizona.EDU,
jtismael@U.Arizona.EDU, kcreath@primenet.com, laj@U.Arizona.EDU,
patrickr@U.Arizona.EDU, rachael@U.Arizona.EDU, sch@U.Arizona.EDU,
shaughan@ns.arizona.edu, tolliver@U.Arizona.EDU, bradt@U.Arizona.EDU

Status: R0

I have some concerns about the epistemic and analysis arguments. In particular it seems that the two epistemic arguments and part of the analysis argument ultimately come back to conceivability arguments. If one is still hesitant about conceivability arguments, then it isn't so clear that the other arguments will take its place. I'll take each in turn.

1) The Argument from Epistemic Asymmetry

The argument is that my knowledge that conscious experience exists derives primarily from my own case. External evidence doesn't play a role in it. (pg 102). This is unlike how we know about other things. From this we can conclude that consciousness can't logically supervene because, "If it were logically supervenient, there would be no such epistemic asymmetry; a logically supervenient property can be detected straightforwardly on the basis of external evidence." (pg 102)

My problem is with the word "straightforwardly." On one interpretation electrons cannot be detected straightforwardly from external evidence. In order to detect electrons you first have to postulate their existence and build a partial theory describing some of their observable properties. Then you have to build some elaborate tools to see one. Why shouldn't I think that we just don't have a partial theory are the right tools to detect consciousness from external evidence. Most of the rest of the mind is an unsolved problem, why not this one?

It seems that my only reason for thinking this can't be done is the conceivability arguments.

2) The Knowledge Argument:

Mary knows everything about the physics/functionality of the mind, but there is something she still doesn't know until she actually sees red.

I have no idea what it would be like to have a complete theory of the physics and functionality of the mind. So my intuitions about Mary are not coming from an understanding that *the* theory of mind doesn't explain consciousness. Rather, it is that I cannot imagine a completed physical/functional theory of mind explaining consciousness. Again it comes to a conceivability argument.

3) From the absence of analysis.

In order for a reductive explanation to work there needs to be an analysis of consciousness whose satisfaction physical facts could imply. But there is no such analysis. Functionalism gives the best hope, but then to explain consciousness is just to explain our ability to manifest some capacity. "But on the face of it, it is entirely conceivable that one could explain all these things without explaining the experience that accompanies the report or the discrimination." (pg 105)

The first part of the absence of analysis argument pretty clearly rests on a conceivability argument. The remainder of the argument is quite interesting and I don't think it is subject to the line I'm running here.

Even if you don't agree with what I've said, there is one point I've used in this argument that I think is important. The mind (even without consciousness) is possibly the most complex thing human's

have ever studied. Vision is probably the most well developed area in the theory of mind and there are still many ongoing debates and just plain holes in that theory. When we imagine a completed functional or physical theory of the mind, I would suggest that we are not even coming close to imagining what such a theory will actually look like. That isn't to say that such imaginings are pointless, but we need to keep the limitations of our imagination in mind.

From chalmers Tue Feb 2 03:36:27 1999
Date: Tue, 2 Feb 1999 03:36:24 -0800 (PST)
From: David Chalmers <chalmers>
To: chalmers
Subject: Epistemic arguments
Status: RO

Thanks to Thony and Josh for their messages. I'm looking forward to more comments and to others' reactions, but in the meantime here are some thoughts (I hope they'll be helpful in the discussion Tuesday). I'll begin with some background and stage-setting, putting the discussion here within the context of the more general metaphysical issues in the seminar.

Modal arguments against materialism, and against other ontological doctrines, take off from the central point that ontological views have modal consequences. For example, if materialism is true, so that in some sense the fundamental constituents of the world are physical, then the physical facts must necessitate all the facts. So one can argue against materialism by arguing that the physical facts don't necessitate all the facts. One can similarly argue for and against other ontological views, by considering the question of whether facts in the (putative) reductive domain (call it the A domain) necessitate facts in the to-be-reduced domain (the B domain).

How does one argue for such a modal claim? This comes down to some deep issues about the epistemology of modality, issues which we'll be concentrating on in this seminar. But the most common way to argue for such a claim is to make an *epistemic* argument. Such an argument tries to establish that there is an epistemic gap between the A domain and the B domain. The thrust of such an argument is usually that there is no a priori entailment from A truths to B truths. From here, it is inferred that there is no necessary entailment from A truths to B truths, so the ontological reduction of B to A can't succeed.

Perhaps the central questions we'll be concentrating on in this seminar is whether the move from an epistemic gap to an ontological gap (from failure of a priori entailment to failure of necessary entailment, for example) is valid, and if so, when and why. But in this section, we're looking at what's involved in the antecedent of such a move -- i.e. the argument for an epistemic gap, which will always be the first step in such an argument. There are a number of ways such arguments can be run. One can use epistemological arguments, arguing directly that knowledge of A truths isn't sufficient to yield knowledge of B truths (even when combined with arbitrary a priori reflection). One can use conceivability arguments, arguing that there is no incoherence in the suppositions that the A truths hold but certain B truths do not. And one can use arguments from conceptual analysis, arguing that B concepts are not of the right type to support a priori entailment from A to B.

The case of consciousness provides some paradigm examples. The question at issue is whether there is an epistemic gap between physical and phenomenal, and in particular whether there is an a

priori entailment from physical to phenomenal. Here arguments from conceivability, epistemology, and conceptual analysis can all be invoked. All of these arguments involve an appeal to some sort of epistemic or conceptual intuition at some point (as do most modal arguments, and indeed most arguments anywhere in philosophy).

I'm not going to try too hard to convince you that there is indeed an epistemic gap in these cases, as we'll mostly be interested in the independent question of whether the epistemic-to-ontic move is valid (a question that remains of general interest for all sorts of issues even if one rejects the epistemic gap in this particular case). But it's good to get a sense of what's involved in supporting such a claim.

Thony and Josh both raise the question: how can we know that there is an epistemic gap between A and B when we don't yet know all the A facts. An anti-materialist will surely concede that we don't yet know all the physical facts, so presumably he or she will have to be making a general claim about what can follow from physical facts, a claim independent of the specifics. I think Jackson intends to make it plausible, for example, that no matter what physical facts Mary knows (even future physical facts), that can't be enough to teach her what red experiences are like. I take it there is at least a strong intuition here (i.e., to the effect that Mary's epistemic gap isn't merely due to her not having the right specific physical knowledge). And I think an opponent would be fighting an uphill battle if they were to argue that with the right specific physical knowledge, Mary really could know what red experiences (and the rest) feel like from within her room. So, we seem to have an intuition of a general epistemic gap here, not just a specific epistemic gap. But it is interesting to think about what might be grounding such an intuition.

One possible general way of grounding the Mary intuition would be: (1) Physical knowledge is all descriptive knowledge; (2) No amount of descriptive knowledge could on its own enable Mary to know what it's like to see red; so (3) No amount of physical knowledge could enable Mary to know what it's like to see red. Of course there are questions about just what "descriptive knowledge" comes to here. One might try to cash it out as "knowledge acquirable in a black-and-white room", or "knowledge acquirable without relevant experiences", or "knowledge about the objective structure of causal systems". For any such reading one could then of course at least argue about whether premise (2) is compelling. My own view is that the general Mary claim has a force which doesn't need to be grounded in such an argument, but such arguments can certainly be used to lend force (and often are, by proponents of epistemic arguments).

Something similar goes for conceivability arguments. When it's claimed that zombies are conceivable, for example, an opponent might respond by saying that we don't know all the relevant physical facts, and therefore don't know just what to conceive of when conceiving of a zombie. In response, the conceivability-arguer might make a general claim such as: for any physical P, it's conceivable that P could be true in the absence of experience. And it might be claimed that this has some strong intuitive plausibility. It seems to many, for example, that the conceivability of zombie scenarios is more or less independent of the specific physical facts in question, and that varying those facts in the conception makes no difference at all to the presence or absence of experience in the conception.

I have some sympathy for this myself, but it's true that such a point

is more compelling if accompanied by some sort of general argument, or some sort of analysis of the underlying relation between physical and phenomenal concepts. For my part, I'd support the claim roughly as follows:

- (1) Physical concepts are all structural-dispositional concepts;
- (2) If B truths are to be entailed a priori by structural-dispositional truths, there must be some analysis of B concepts in structural-dispositional terms;
- (3) There is no analysis of phenomenal concepts in structural-dispositional terms; so
- (4) Phenomenal truths are not entailed a priori by physical truths.

In a way, this corresponds to the point stressed in the book that the argument from the absence of analysis is what lies at the root of the epistemic arguments; and premise 2 in particular corresponds to the point that "structure and function only ever adds up to more structure and function".

Of course, an opponent could question premise 3. (They might also question premises 1 or 2, but I think this would probably not be fruitful in the end.) I take it that a dispositional (functional) analysis would be the only candidate of any promise, and that ultimately there are only a few promising dispositions in the vicinity. Here, though, note that the epistemic arguer can mount a case against such analyses without needing to worry about the precise nature of specific physical facts. The claim will just be that for any disposition in the vicinity, there is an epistemic gap between that disposition and experience. For example, for any such disposition, it's conceivable that it be present without consciousness; or one could know all the facts about such dispositions without knowing all the facts about consciousness. Someone might still question this, but at least we have narrowed down the scope of the discussion from worries about future physical facts.

All this bears directly on Josh's question about priority between arguments from conceivability, epistemology, and analysis. To my mind, the question of priority is subtle. I think that each of the claims have a certain *prima facie* force that needn't be grounded in the others, but that underneath there is something of a relation of mutual support going on. It would be surprising to me if the conceivability argument were seen to ultimately carry the burden, as there are a number of people who have the Mary intuition and the analysis intuition while being unsure about or even denying the full conceivability intuition. But no doubt there is something of a circle of mutual support here, and it is interesting to uncover its structure.

If I were to try to articulate the substratum underneath the intuitions, I guess I would hold as above that the full conceivability claim is grounded in the deeper point about absence of the relevant sort of analysis, i.e. in a claim about the nature of phenomenal concepts, and it may be that the epistemological arguments are too (though I think that both of these arguments can have *prima facie* force to one who is uncommitted on the abstract underlying point). It's arguable in turn that the point about analysis is itself grounded to some extent in intuitions about conceivability and epistemology, but note that these are now much more limited claims than the full-blown claims (concerning arbitrary physical facts) which we

started with.

I also think, though, that the argument from absence of analysis has independent support, e.g. in the sort of reasoning which holds that "for any function, we can always raise the question, why is the performance of this function accompanied by experience, and this will always be a nontrivial further question?". This sort of further-question reasoning (which I rely on most in my paper "Facing Up the the Problem of Consciousness") can arguably be supported by conceivability and epistemological considerations, but I'd argue that it has intuitive support which is more or less independent of and prior to such considerations. So if I were to rest on any point as supporting the central burden, it might be on points like this concerning the conceptual distance (at least in a priori space) between functional and phenomenal concepts. But I think it is most accurate to see things in terms of a complex web of support among a number of related considerations.

A few specific points re Thony and Josh.

Re the "Dark Room":

- >1. Imagine Mary knows all there is to know about light.
- >2. But Mary is in a completely dark room.
- >3. Suppose Mary is swinging a bar magnetic abover her head in the dark
- >room. Nothing happens. Then someone turns on a light.
- >4. She's learned something new: light isn't electromagnetic wave
- >propagation after all.

I presume that premise 1 is meant to say that Mary knows everything *physical* about light (in order to engage appropriately with 4, and to provide an appropriate analogy). Here I'd want to know in what intuitive sense "Mary has learned something new". It seems to me that intuitively, the only new thing Mary has learned is what it is like to see light; i.e. some facts about light *experience*. From here, one might well infer that facts about light experience are not reducible to facts about EM wave propagation. But it's hard to see a more general argument about non-experiential aspects of light getting off the ground; so it's hard to see that we have any analogous plausible but fallacious argument in the vicinity. At least, I'd need to see a lot more spelling out of just what is supposed to be going on in this scenario, and what is supposed to be learned, from the point of view even of the 18th century scientist.

Re: the epistemic asymmetry argument, and "a logically supervenient property can be detected straightforwardly on the basis of external evidence."

This was probably too quick on my part. Properties such as "had meat for breakfast in 10,000 BC" are presumably logically supervenient but not straightforwardly detectable. On the other hand, *given* relevant physical and microphysical knowledge, such properties are fairly straightforwardly detectable. I guess something similar would apply to electrons. They're detectable only nonstraightforwardly from external evidence, but determinable pretty straightforwardly from complete microphysical knowledge. Whereas even complete microphysical knowledge leaves a gap in the consciousness case. Of course an opponent might deny the epistemic intuition here, and say that with the right physical knowledge, one's knowledge of consciousness would be fairly straightforwardly derivable. I agree that this epistemic asymmetry argument is pretty close to both the conceivability and direct epistemological arguments, which is why I don't rely on it

much.

Re the knowledge argument: I think I'd want to deny that this is a conceivability argument. The argument isn't that one can't conceive of Mary knowing what it's like to see red, it's that we have good reason to believe that she won't know what it's like to see red. See above.

Re the analysis argument: I've tried to spell out the extent to which this does and doesn't rely on conceivability considerations.

Re the brain being complex: Of course this is so. But still, a physical theory of the brain's functioning has a certain shape, irrespective of the details, and it may be that there are things we can know about what theories of such a shape can and can't explain, by virtue of certain very general principles that are independent of the details. (Think about someone arguing that no theory of pure statics [without a temporal element] could explain dynamics, for example.) E.g., if it were the case that (1) a physical theory can only explain structure and function, and (2) to explain structure and function is not to explain experience, then one could draw conclusions about the limitations of physical theory in explaining experience. Of course work is being done by general premises such as (1) and (2), which can be questioned, but the point is that we can do a good amount of work in assessing such premises even in advance of knowing the specific details of physical theories of the brain.

I certainly don't want to say that empirical knowledge of the brain is completely irrelevant in these matters; and it's certainly vital for understanding all sorts of philosophically important things about the mind. But there is a lot of philosophy we can do in advance of understanding all the details. One had better be sure that the philosophy one does is compatible with any details that emerge, of course. My own view is that attention to the details of neurophysiology and other empirical areas tend to give the epistemic considerations in question fairly strong support. One might reasonably disagree about that, or hold that some future neurophysiology will show us something pretty different, but I suspect that for such a view to have much weight, it's going to have to be supported by a good amount of relatively a priori argument on the conceptual and epistemic matters in the vicinity.

These are all deep issues that deserve attention, and I'd be interested to hear any thoughts.

--Dave.

From bradt@u.arizona.edu Wed Feb 3 01:29:52 1999
Date: Wed, 03 Feb 1999 02:23:55 -0700
From: bradt <bradt@u.arizona.edu>
X-Accept-Language: en
To: bradt@u.arizona.edu
Subject: minutes from seminar meeting
Status: RO

Hi all:

Here is a summary of our discussion Tuesday--feel free to add anything that I may have left out.

First, questions about the unidirectionality of the supervenience relation were addressed, relating to some of our earlier online

discussion.

We also briefly touched on the definition of materialism and how it is that the logical possibility of zombies would entail the falsity of materialism. This lead us into a discussion of the relationship between conceivability and possibility. It was pointed out that the relevant notion of conceivability has to be stated carefully, so as to rule-out the conceivability of Goldbach's conjecture being false if in fact it is true, for example. Some other remarks were made to the effect that the close relationship between conceivability and possibility is (perhaps pragmatically?) desirable, since any gap between conceivability and possibility might lead to mysteriously inaccessible possibilities or impossibilities.

Everyone seemed to be clear about supervenience, possibility, and necessity. We spent most of our time discussing epistemic arguments against materialism, especially the case of what Mary didn't know. Thony was asked to restate the issues he had raised in online discussion. We then debated some of the points raised there. In particular, we discussed whether or not epistemic arguments were weakened by the fact that the physical facts which are claimed to be insufficient for knowledge of phenomenal facts are left unspecified. Some claimed that this did make epistemic arguments weak. Others were essentially in agreement with Dave that it was sufficient to notice that the relevant physical facts would merely provide more information about structure and dynamics, and that such information would always fall short of providing knowledge of the phenomenal facts.

We also ended up discussing subjectivity and objectivity, as well as concepts (and whether or not they had both subjective and objective aspects to them). Somehow (!) this also lead us into discussing Wittgenstein on private language and on the beetle in the box, and Dennett with regard to purported differences that "don't make a difference" (such as Orwellian vs. Stalinesque revisions).

Brad

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 3 06:02:08 1999
Date: Wed, 3 Feb 1999 06:02:04 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Perplexity
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Thanks to Brad for the minutes. I would be interested to see more on these points.

I'm a little perplexed that so little is being posted to the mailing list. Only a couple of messages in the last 6 or 7 days, even though I asked that everybody post something about the current round of readings by now. If this goes on, we are going to fall a long way behind. Is there a problem? If so, please let me know (private e-mail is fine). Any suggestions as to methodology are welcome.

As things stand, I think I have to be more explicit. There is a weekly writing assignment, which involves posting at least a page or so of questions and/or comments on the current round of readings (and feel free to post more). Students are also expected to engage in a good amount of mutual discussion of each others' contributions, in

addition to this (N.B. "in addition", i.e., the initial writing assignment shouldn't be primarily reactive). These writing assignments are a core part of the seminar.

I know the online seminar is an unusual arrangement and is not ideal, and I hope we'll be meeting as a normal seminar soon (no news on the visa yet, alas). In the meantime, though, I think some effort at participation on everyone's part can make this a fruitful and productive arrangement. Again, any suggestions are welcome.

I'll look forward to seeing everyone's questions and comments on the second set of readings in the next day or so, followed by a good round of discussion. You should also be reading over Naming and Necessity for the third round; your thoughts on that will be due early next week. Although we won't be focusing on the 2-D material until the fourth round, I suggest that those of you who have read N&N already read over the 2-D material first (esp. in TCM and The Components of Content), and then read over N&N with that material in mind, trying to understand it in those terms. I'll give advance notice that at some point fairly soon there will be an assignment asking you to translate all or most of the central examples in N&N into the 2-D framework. That should help you get a working understand of what the 2-D framework is all about and of how to use it, which will be important for the rest of the seminar.

--Dave.

P.S. We're now ready to switch to the listserv, so all future contributions should go to modality@listserv.arizona.edu.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 3 09:33:42 1999
Date: Wed, 3 Feb 1999 10:33:00 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: On Mary and Experience and Concepts
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

All, In reply to Dave's request for more info on Tuesday's discussion, let me attempt to elaborate the line that Thony and I were attempting to develop. (Though don't hold Thony accountable for what follows!) I don't think that the argument would convince anyone not already convinced that Mary can know what it's like to see red, but perhaps it focuses the argument on the issues on which it should be focused: the nature of *concepts* about the phenomenal.

(1) Assume that Mary is not only an expert neuroscientist, but an expert developmental psychologist as well. She knows how infants develop all the concepts that they do on the basis of their innate mental structure and the sensory input that they receive. Not a contentious assumption, I hold.

(2) Assume that infants develop all concepts on the basis of phenomenal states and their innately given structure. This is controversial. The idea is that sensory input leads to phenomenal states in infants, and they form their concepts on the basis of those phenomenal states (probably together with the feedback they get from action - which we could regard as simply more sensory input). Note: I don't claim that this is the only way in which one could develop concepts. Surely you could (logically) build a machine that develops concepts on the basis of input without going via the sensory states that that input generates, but it does not *seem* as though

this is how we do it.

(3) Mary has the innate mental structure that infants do.

(4) Mary has the concepts that, say, the normal adult human being has. In particular, she has the concept red. After all, she knows everything that there is to know about colors, and to do this she must have all the color concepts.

(5) So, Mary will be able to subtract from her concept of red the innate structure that went into it, and come up with the phenomenal experience on the basis of which it was constructed. In other words, she can simply reverse-engineer her concept of red, and do in reverse what the infant does in forward. Thus, Mary-the-developmentalists does know what it's like to see red.

Here's the predicted reply from Jackson and friends: (4) is false. Mary does not have the concept of red. Or, perhaps more accurately, she does not have the concept of red that is important for this discussion. She may have *a* concept of red in the sense that a blind person can have a concept of red (it's a color, people see it, tomatoes are red, as are fire engines, etc.) but she does not have the phenomenal concept of red, of which red qualia is a *constituent.* You cannot have the phenomenal concept of red without qualia because the red quale is a (proper?) part of it. Knowing-that can include phenomenal states, because it includes phenomenal concepts, and phenomenal states are part of (some) phenomenal concepts. And the only way you can generate these concepts is by having certain selective types of input from the environment. Mary is in her dark room (or is color blind), and thus cannot get (or operate on) this type of input. (And she cannot provide this input endogenously either - she cannot visualize in red.) Hence she could never concept of red from which to sift out or distill the phenomenal experience of red.

As I say, this is a fair reply on the part of Jackson. (And actually, there are other problems with the argument too.) It seems to me that the essential issue is now the nature of phenomenal concepts. If we accept the claim that qualia are constituents of phenomenal concepts, then we can say that there are two ways one can come to know a phenomenal property: via the experience itself, or via the phenomenal concept. Indeed, we might even be so bold as to suggest that qualia simply *are* phenomenal concepts. Perhaps qualia can have inferential roles, enter into reasons, and do some of the other nifty things that concepts do. Actually, it occurs to me that this is Lycan's line on qualia (I think): they are concepts that are only intra-subjectively potent; they are, so to speak, words in the experiencers language of experience, that can't make the leap into public discourse. (But perhaps I have Lycan wrong.)

What got me thinking along these lines was the attempt to bridge the gap between the objective (concepts - available for intersubjective discourse, the box a la Wittgenstein) and the subjective (qualia - not available for intersubjective discourse - the beetle). Now: if infants can build up from the subjective to the objective, then why can't Mary go in the other direction. Of course, Fodor's reply is that infants don't build up from the subjective to the objective: all concepts are innate. But as long as you're with Auntie on this one and think that Fodor's wrong, maybe there is something to the thought that Mary can do in reverse what we - as infants - do forward.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 3 10:27:49 1999
x-sender: agillies@pop.u.arizona.edu
Date: Wed, 3 Feb 1999 11:31:02 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: On Mary and Experience and Concepts
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

>Here's the predicted reply from Jackson and friends: (4) is false. Mary
>does not have the concept of red. Or, perhaps more accurately, she does
>not have the concept of red that is important for this discussion. She may
>have *a* concept of red in the sense that a blind person can have a
>concept of red (its a color, people see it, tomatoes are red, as are fire
>engines, etc.) but she does not have the phenomenal concept of red, of
>which red qualia is a *constituent.* You cannot have the phenomenal
>concept of red without qualia because the red quale is a (proper?) part of
>it. Knowing-that can include phenomenal states, because it includes
>phenomenal concepts, and phenomenal states are part of (some) phenomenal
>concepts. And the only way you can generate these concept is by having
>certain selective types of input from the environment. Mary is in her dark
>room (or is color blind), and thus cannot get (or operate on) this type
>of input. (And she cannot provide this input endogenously either - she
>cannot visualize in red.) Hence she could never concept of red from which
>to sift out or distill the phenomenal experience of red.

I think Tim is right in predicting this reply. One worry: if we say (in the premises) that Mary doesn't have the full concept of red but just the non-phenomenal component, then aren't we begging the question against the materialist? After all, the *conclusion* is supposed to be something along those lines. On the other hand, the materialist's use of premise (4) (i.e., that Mary has the full concept of red) seems to beg things the other way; so this might be an indication not that this Tim's argument is an argument *for* materialism, but rather an argument that the Mary-the-neuroscientist-case isn't a good argument *against* materialism.

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 3 10:43:08 1999
Date: Wed, 3 Feb 1999 11:32:26 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Re: Perplexity
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I have a question regarding Thony's example of the government or

"body-politic" (I can't remember the precise phrase now). Did you mean that, even in principle, there are no illuminating accounts of a political process in terms of basic physics. If you meant in principle, I guess I'm confused. In practice, explaining a high-level phenomenon like a political process might be very cumbersome using physics, but it seems like, if that political organization you spoke of is entirely physical, there is nothing precluding such an explanation in principle. "not very illuminating" is perplexing in it's own right. What makes something "illuminating"? That's a strange notion once you've conceded that the process referred to is nothing over and above physics. I guess what I'm asking is, what is the role--ontologically, functionally, explanatorily or whatever--of concepts in explanations and why should they be NECESSARY if a supervenience relation holds between the phenomenon and a set of entirely physical facts? For that matter, WHAT IS A CONCEPT in this sense that we keep using it?

Erik L.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 3 12:24:02 1999

Date: Wed, 3 Feb 1999 13:27:42 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Timothy J Bayne <bayne@U.ARIZONA.EDU>

Subject: Conceivability and Knowledge arguments

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Dave suggests that conceivability arguments are distinct from knowledge arguments. The Mary argument is a knowledge argument: Mary knows all the physical facts (A facts) but she doesn't know all the facts about experience (B facts), therefore Zombie arguments are conceivability arguments: I can conceive of my zombie, i.e. of a being with all my physical properties but none of my phenomenal properties, therefore . . .

I'm not so sure that these arguments are really distinct at any deep level of analysis (although there's a caveat below). First, note that one has to conceive of Mary and locate her in the distant future of neuroscience. What does this involve? Well, it's hard to say. It might involve visualizing a certain scene (say, Mary wearing a white lab coat), and labelling it *Mary the expert neuroscientist who knows everything about the brain*. What does conceiving of my zombie involve? Well, perhaps something similar. I have a picture of someone just like me, but there's nothing it's like to be him - nobody's home and the lights aren't on - and I stick a label on him: *My zombie twin.*

But in fact, I doubt that we do anything *like* this in either case. Conceivability does *not* seem to involve the imagination - at least that's what it seems like to me in *these* cases. It's *not* like trying to work out in your head whether you can fit a ball of diameter two inches inside a triangle of sides 3, 4 and 4 inches which has an outer perimeter of 1/2 an inch thick.

So what is going on in these cases? Well, it seems to me that in both cases you have the following intuitions or you don't.

(a) Necessarily, all physical properties are functional or structural properties, and phenomenal properties aren't functional/ structural properties; OR, the very similar intuition: (b) Nec., all physical facts (or, perhaps, propositions) are exclusively composed out of functional /structural concepts, and phenomenal concepts are not structural/functional concepts.

(a) is at work in the zombie cases, (b) is at work in the knowledge cases. If there is a one-one correspondence between facts and properties as is commonly supposed in setting these cases up, then it is hard to see why one should have (a) and not (b), or vice-versa. (Caveat: Of course, some do reject the claim that if you've got two facts you've also got two properties, but then you can accept that Mary does learn something new without rejecting physicalism.)

But imagination does seem to play some role in conceivability. Consider the following problems: (1) is logically it possible to be an agent without, being a perceiver (or vice-versa)? (2) Is it possible for one subject of experience to have two distinct (spatially separated) bodies at the same time, so that, e.g., one of my bodies might be in Tucson, and the other one in Texas at the same time, and they are simultaneously conscious? Well, people seem to have very different intuitions about both of these cases? Why? Are they: (1) actually imagining different scenarios? or (2) imagining the same sorts of scenarios, but putting different labels on them because of the differing structure of their concepts? I dunno, but I'm tempted towards (2).

Perhaps conceivability and knowledge arguments are different types of arguments because they are different ways of pumping the same intuitions (i.e. the same concepts), but I'm not even sure that this is true.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 4 03:31:18 1999
Date: Thu, 4 Feb 1999 03:31:14 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: On Mary and Experience and Concepts
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Tim's and Thony's line of argument is interesting. Speaking as someone sympathetic with Jackson, I'd be tempted to endorse the predicted reply to Tim's argument, contesting (4). Actually, I'd be tempted by a somewhat fuller two-part reply.

(1) There are certain concepts of "red" that Mary has in the black-and-white room, but these are not concepts that require the experience of red for their possession. They might be concepts analyzable as "experiences of the sort caused in normal observers by certain paradigm objects (the objects which normal observers label 'red') (for her concept "red experience") or "objects of the sort that produce red experiences (i.e. the same sort of experiences as are produced by the paradigm objects) in normal observers under normal conditions" (for her concept "red object"). It's plausible that the "red" concepts that Mary possesses have this sort of nature, but I think it's clear that neither of these concepts requires Mary to have

ever had a red experience, and there's no reason to believe that possession of either concept will allow her to know what a red experience is like.

(2) There are other concepts of "red" that Mary doesn't have in the room, but acquires later on having her first red experience. Such a concept might include the concept R corresponding to her new belief, tomatoes look R. (Of course she already knew they look red, in the above senses.) Mary doesn't have this concept in the room, so she won't be able to use it to figure out what red experiences are like.

I think that (1) alone is a sufficient reply to the argument, so the anti-materialist doesn't need to put special weight on (2). But I think that (2) is plausible all the same. I think it's more or less common sense to say that a blind person can have only an "indirect" concept of red experiences, as opposed to the "direct" concept that is possessed by someone who has actually had the experiences.

Is this question-begging? I hope not. (1) seems more or less like common sense to me, and (2) isn't really going out on a limb. It seems to me that both of these are claims that even many materialists might endorse. One needn't even rely on claims about what Mary could and couldn't figure out in principle; one can just think about the ordinary concepts possessed by a non-omniscient colorblind Mary, and the concepts possessed by an ordinary person who can see colors.

(Incidentally, "begging the question" is a very interesting phenomenon in its own right. Almost any philosophical argument can be accused of begging the question. Say it is a valid argument from premise P to conclusion not-C, arguing against a C-ist. Then it's obviously equally the case that if C is true, then P is false. So clearly on the C-ist position, not-P. So a C-ist can accuse the original arguer of begging the question against them by asserting P. Obviously this can't be quite right. I think the most important distinction between question-begging and non-question-begging arguments is whether the premises have plausible independent support antecedently to considering the conclusion.)

In any case Tim is right that these issues about phenomenal concepts are very important. I try to develop an account of phenomenal concepts of the sort discussed in (2), on which the concept R is partly constituted by the red experiential quality itself, in my "The Content and Epistemology of Phenomenal Belief" (online as the third Princeton talk).

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 4 23:29:53 1999
X-Accept-Language: en
Date: Fri, 5 Feb 1999 00:12:29 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Brad Thompson <bradt@U.ARIZONA.EDU>
Subject: Our ignorance of "all the physical facts"
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

It seems to me that the most defensible materialist response to the anti-materialist arguments we've been discussing and reading about is the one that the Churchlands and Dennett have given (I think all 3 philosophers have given this basic response to Jackson's argument, Nagel's argument, and the conceivability of zombies). The response is that we don't know

"all the physical facts", so that what we *succeed* in imagining when we try to imagine Mary's situation (or similarly being a bat, or a physical duplicate of me without consciousness) is not what the thought experiment *asks* us to imagine. The materialist assumption that goes with this reply is that *if* we did know all the relevant physical facts, zombies would be inconceivable (and we'd know what it is like to be a bat, and we'd be able to imagine being in Mary's situation and wouldn't conclude that she didn't know what seeing red was like).

Dave's response to this objection appears to be that even though we do not know the specific physical facts needed, we know what *types* of facts that additional knowledge would be--more structure and dynamics. And adding more structure and dynamics would still not give us a reductive explanation. Look at functionalist theories of qualia. They radically fail to give an account of why a red experience has the particular quale that it does, or any quale at all. The inverted spectrum argument shows this, but the shortcomings of functionalist accounts doesn't really require argument--they simply fail to provide any explanation for why there are qualia or for why mental states have the particular qualia that they do. But in the case of functionalism, I think it is pretty easy to see that adding extra functional information could not possibly remedy the situation.

But what about some yet unknown neurophysiological information--could this close the epistemic gap (is this what Mary knows, that we don't)? Well, it is difficult to imagine how such a view could resist positing what Dennett would call "wonder tissue". What non-functional property does the brain possess which makes it conscious?

Though I think that the above considerations help against the materialist response I've been discussing, I think we have to concede that the anti materialist arguments are weaker than they would be if we did in fact have total physical knowledge. The materialist might insist that this concession entails that the anti materialist arguments simply fail. But it seems to me that they should not be too comfortable in resting their own view on a check which, *prima facie*, they probably cannot cash. One can view the anti materialist thought experiments as providing the *prima facie* evidence that that check won't be cashed. Perhaps we are left then with the question of where the burden of proof lies--on the materialist or on those who deny materialism.

PS-- On another note, there is a wonderful example of what Dave was talking about with regard to begging the question. In Philosophical Naturalism, David Papineau gives an argument for materialism which has the falsity of epiphenomenalism as one of its premises. Jackson, on the other hand, argued against materialism and then uses that to argue for epiphenomenalism.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 5 10:12:42 1999
Date: Fri, 5 Feb 1999 10:59:06 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: Re: Our ignorance of "all the physical facts"
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

In response to Brad.

> Dave's response to this objection appears to be that even though we do
> not know the specific physical facts needed, we know what *types* of
> facts that additional knowledge would be--more structure and dynamics.

> And adding more structure and dynamics would still not give us a
> reductive explanation.

What is it that 'structural' means here? Does it mean physical structure or is it more general? If it is more general, then I'm not sure why consciousness isn't structured.

One note on the materialist insufficient-info rebuttal. We have been discussing the rebuttal as if it has the following form.

1. We don't know all the physical facts.
2. You can't have a good conception of Mary (zombies etc.) unless you know all the physical facts.
- C. Therefore the arguments fail.

I think an important step is being left out of the real argument though.

3. Everything other than consciousness seems to be explainable in terms of physics
4. It would be weird if there were only a single phenomena that couldn't be explained by physics.
- C. Therefore I have better reason for thinking I'm missing something than for thinking physics can't explain qualia.

You can, of course, argue whether you actually have better reason for this. But its important to see that the materialist is comparing possibilities and judging one to be more plausible.

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 5 19:42:54 1999
Date: Fri, 5 Feb 1999 19:42:50 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Our ignorance of "all the physical facts"
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I'm heading out of town for a couple of days, but a quick reply to Brad and Josh.

Brad says:

>It seems to me that the most defensible materialist response to the anti
>materialist arguments we've been discussing and reading about is the one
>that the Churchlands and Dennett have given (I think all 3 philosophers
>have given this basic response to Jackson's argument, Nagel's argument,
>and the conceivability of zombies). The response is that we don't know
>"all the physical facts", so that what we *succeed* in imagining when
>we try to imagine Mary's situation (or similarly being a bat, or a
>physical duplicate of me without consciousness) is not what the thought
>experiment *asks* us to imagine. The materialist assumption that goes
>with this reply is that *if* we did know all the relevant physical
>facts, zombies would be inconceivable (and we'd know what it is like to
>be a bat, and we'd be able to imagine being in Mary's situation and
>wouldn't conclude that she didn't know what seeing red was like).

Actually, I don't think this is really Dennett's main line of response (though I know he says something like this somewhere), and for good reason. Dennett holds that there is nothing more to qualia than certain reactive dispositions -- i.e. he holds that qualia are

functionally definable, and/or that any sense in which qualia aren't functional concepts is a sense in which qualia don't exist. It follows from this that on Dennett's view, once we know the appropriate facts about which functions are performed -- the facts about discrimination, integration, higher-order monitoring, control of behavior, and verbal report -- we know everything. So Mary knows everything as soon as she knows that much (the apparent residue is just an illusion), and it's not really coherent to suppose that something could have that functional profile without having qualia (or: any sense of "qualia" in which this is coherent is a sense in which we don't have qualia). So relying on further unknown physical facts is more or less irrelevant to Dennett's position.

I respect this position, as I think it is the most coherent line for a materialist to take. If qualia are to be conceptually entailed by physical facts, it will be by virtue of functional analysis of phenomenal concepts, so what matters are the functions, not the underlying physical facts. And the relevant work in seeing how the entailment can go through is in a sense conceptual (an a priori entailment will be grounded in an a priori (if deep) conceptual analysis, so one doesn't need to wave one's hands to future physical discoveries). Someone like Lewis is also admirably clear about this.

Something like this comes out in what Brad notes:

>But what about some yet unknown neurophysiological information--could
>this close the epistemic gap (is this what Mary knows, that we don't)?
>Well, it is difficult to imagine how such a view could resist positing
>what Dennett would call "wonder tissue". What non-functional property
>does the brain possess which makes it conscious?

The obvious trouble is that the explanatory relevance of neurophysiology is always by virtue of the functions that the neurophysiology performs. So it's not the neurophysiology itself which is crucial to crossing the gap, it's the function, and we're back to the previous sort of position. To suppose that there's some other way that the neurophysiology might be relevant seems to lead to "wonder tissue" -- perhaps conceptual entailment by virtue of intrinsic properties rather than by virtue of functional role? -- and doesn't seem to hold up to a close analysis of what neurophysiology is all about. (Actually, I'm not completely hostile to the idea that there could be some sort of intrinsic entailment, but taken seriously the idea leads straight to "panprotopsychism").

Of course the "we don't know all the physical facts yet" position is very tempting at first blush, but it's not clear to me that this temptation should survive a careful analysis, and whether one can make coherent sense of the hopeful gesture. Entailment by virtue of a functional analysis I can understand, but then the real burden will be carried by analysis, not by the empirical details. And it's not at all clear how any other sort of entailment from physical to phenomenal could work, given the nature of physical concepts (and in any case, the central work will still be done by an analysis, which is conceptual).

Of course a materialist might say that future physical discovery will lead to something utterly unlike the kind of physical knowledge we currently have, and which won't be subject to the same constraints. I think of this as more of a "new physics" than a "new neurophysiology" approach, and certainly not the sort of thing a Churchland or Dennett would be tempted by (it would require the sort of "real humdinger"

than Churchland derides). Even here, I think that close analysis reveals that the options are pretty limited.

All this being said, if someone wants to try and develop the "future physical knowledge" line into some sort of coherent position that evades the obvious objections, I think that would be a very useful service indeed, and I'd be very interested to see it.

Josh writes:

>What is it that 'structural' means here? Does it mean physical
>structure or is it more general? If it is more general, then I'm not
>sure why consciousness isn't structured.

Well, consciousness certainly has structure but (a) it's structural properties don't seem to exhaust the explananda (which most centrally involve intrinsic non-structural features), and (b) the structure it has doesn't seem to be the right sort to hook into entailment by physical facts. As you suggest here, the structure isn't physical structure. Physical structure seems to involve (a) spatiotemporal relations and (b) relations in certain underlying physical spaces, such as Hilbert spaces and the like. But this sort of structure is clearly conceptually compatible with the absence of consciousness. Even if one could get some sort of structural match-up (e.g. one might suppose that *given* that we have experience, one will have some sort of structural coherence), it's even less clear how one could exhaustively analyze the concept of experience in terms of static physical-structural concepts than in terms of functional concepts.

(But still some interesting conceptual space to play with, so one might like to try!)

>One note on the materialist insufficient-info rebuttal. We have been
>discussing the rebuttal as if it has the following form.

>
>1. We don't know all the physical facts.
>2. You can't have a good conception of Mary (zombies etc.) unless you
>know all the physical facts.
>C. Therefore the arguments fail.

>
>I think an important step is being left out of the real argument
>though.

>
>3. Everything other than consciousness seems to be explainable in
>terms of physics
>4. It would be weird if there were only a single phenomena that
>couldn't be explained by physics.
>C. Therefore I have better reason for thinking I'm missing something
>than for thinking physics can't explain qualia.

>
>You can, of course, argue whether you actually have better reason for
>this. But its important to see that the materialist is comparing
>possibilities and judging one to be more plausible.

Well, of course there are lots of good reasons to be a materialist. That's more or less taken for granted in these discussions, by virtue of the fact that materialism is always regarded as the plausible default position which has to be dislodged, rather than vice versa. But of course the materialist still has to answer counterarguments. And as a counterargument, "maybe we're missing something" is not the strongest. Of course things would be different if one could provide some sort of faint gesture to just what we might be missing (not

asking for an explanation, just for the conceptual room for an explanation).

The anti-materialist can also try to "explain away" the force of (3) by noting that all those "everything else"s that have been explained seem to come down to the explanation of physical structures and of functions, and quite obviously so, whereas experience is *prima facie* not a matter of explaining either. So one has grounds right there for not giving the inductive argument too much force.

I'm not really trying to convert anyone to anti-materialism here, and there are all sorts of reasons why materialism is an attractive position. But I do think the materialist has to face up to bullets that need to be bitten. As I said, I think the most coherent sort of materialism is the sort that just denies the epistemic intuitions, and is either functional or eliminative about the concept of qualia. Of course what the position gains in coherence it may lose in plausibility. There are other intuitively "more plausible" positions out there at least at first blush, but the question is whether they are philosophically coherent. Hopefully the rest of the course will provide us with some of the tools to examine such positions in some depth.

--Dave.

P.S. Kripke questions and comments are due by Tuesday. In the meantime this discussion is interesting, so please go ahead with any further thoughts in this vicinity or elsewhere re the second set of readings. A number of people still owe comments re this week -- soon, please!

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 5 01:35:36 1999
Date: Fri, 5 Feb 1999 02:24:54 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik A Herman <erikh@U.ARIZONA.EDU>
Subject: Re: On Mary and Experience and Concepts
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I agree with Jackson but I don't find his examples very compelling. Learning from feeling and learning from being told/reading it are two totally different ways to experience something. They are different in kind and so can't really be compared. It's like telling someone blue feels cold-- you aren't gonna get it unless there's a one to one correspondence between blue and cold or sight categories and olfactory categories in general. If there WERE this correspondence then I would say that it's fair to translate visual information into verbal information in which case I think we WOULD be able to know red2.

The situation with Mary: Again, I think she certainly will learn something when she sees the color TV. But what won't be immediately obvious to her is how the colors correspond to her prior interpretation of all the black and white things, other than there just being differences.

As for the Modality Argument, I am concerned that there are inherent lawlike connections among physical things, and perhaps consciousness is one of these-- that by virtue of having a brain physically identical to mine, a duplicate must be conscious and in the same conscious state.

Oh, not to jump around but back to Fred, does it ruin everything if qualia is unique to each individual? Could it be that what you feel is your

relation to that thing (so it involves both partys to the interaction)?

Also, I can't help but think that Mary would dream in color-- does this count as a qualia?

Erik H.

From owner-modality@LISTSERV.ARIZONA.EDU Sat Feb 6 16:01:18 1999
Date: Sat, 6 Feb 1999 16:50:35 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: Re: Perplexity
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Here are a few general questions about last weeks material. A bit of a regression in the discussion. I'm afraid, but here goes...

In section IV of *Epiphenomenal Qualia* Jackson discusses three arguments that might be presented to show that qualia might have some role in determining brain states. I suppose that his commitment to the impossibility of qualia having any causal efficacy stems from the rest of his argument. If Mary were to know everything about physics, neurophysiology and so forth, but had not had the experience of seeing red, and seeing red had some causal role in determining brain states, then, presumably, she would not have all of the physical knowledge there is to have. There are two points I am not clear about here:

First, I am a bit confused about the phenomenal/ psychological division. It seems to me that the phenomenal quality of our experiences does play a causal role in determining behavior. Isn't it the feeling of pleasure that one has when one sees a particularly beautiful sunset that leads one to try to have such experiences? I think this is probably a misunderstanding on my part, but it does seem that when I want to eat a pint of Ben and Jerry's, it is the phenomenal quality of the experience that I find particularly appealing. How does this work if we speak in purely psychological terms?

This also gave me a worry in regard to the zombie argument. We are asked to admit the logical possibility of a world in which there are beings physically and functionally identical to us, but without consciousness. But why would these beings look at sunsets or eat ice cream? I don't know how behavior such as we exhibit is to be explained in terms of only functional states.

Secondly, is it a fact that prior to having a particular kind of sensory experience we are incapable of having experiences of that sort? Could Mary, having read about the experience of seeing red and having examined various MRIs and what not of people as they experience red, devise some sort of device to stimulate her brain in such a way as to have the phenomenal experience of red?

Sorry to be rambling somewhat. A somewhat unrelated question... In the first part of chapter you talk about the form of dualism you are advancing. You suggest that phenomenal properties are basic and that there must be fundamental psychophysical laws relating these phenomenal properties to physical properties. Will these laws relate particular physical (brain) states to certain conscious experiences? Is the point of this that it is a contingent matter what psychophysical laws actually exist?

Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Sun Feb 7 23:58:47 1999
Date: Sun, 7 Feb 1999 23:58:43 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Qualia
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Nobody needs to worry about "retarding discussion". It's important to be clear about things, and we can easily have numerous discussion topics open at the same time, so it's OK for a discussion of one's weeks topics to drag over into the next week. As long as the next week's discussion goes ahead just the same, things should be fine (N.B. everyone should make their initial contribution near the beginning of the week, though, to maximize discussion potential).

Here are some random notes on comments, mostly about qualia, by Erik H. and Anthony.

Erik writes:

>I agree with Jackson but I don't find his examples very compelling.
>Learning from feeling and learning from being told/reading it are two
>totally different ways to experience something. They are different in
>kind and so can't really be compared. It's like telling someone blue
>feels cold-- you aren't gonna get it unless there's a one to one
>correspondence between blue and cold or sight categories and olfactory
>categories in general. If there WERE this correspondence then I would say
>that it's fair to translate visual information into verbal information in
>which case I think we WOULD be able to know red2.

Well, for our purposes, Jackson is just trying to argue that the phenomenal facts are not entailed a priori by the physical facts. As far I can tell, you're agreeing with this here (e.g. you agree that Mary gains new knowledge that complete knowledge of physics and physiology can't give her). Of course it's then a further question whether this epistemic gap leads to an ontological gap, and that's one of the things we'll be looking at a lot more deeply.

>The situation with Mary: Again, I think she certainly will learn
>something when she sees the color TV. But what won't be immediately
>obvious to her is how the colors correspond to her prior interpretation of
>all the black and white things, other than there just being differences.

Right, that may not be obvious from introspection. On the other hand, Dennett has a nice passage where he suggests that Mary might be able to figure out which color is which, etc, because she might know that different colors will cause slightly different reactions (e.g. emotions, twinges, flickers, etc) in her which she can identify. So when the scientist's try to trick her by presenting her with a blue banana, she isn't fooled into thinking it's yellow. I think that's entirely possible (it's an empirically open question), though I don't think it impacts the materialism debate.

>As for the Modality Argument, I am concerned that there are inherent
>lawlike connections among physical things, and perhaps consciousness is
>one of these-- that by virtue of having a brain physically identical to
>mine, a duplicate must be conscious and in the same conscious state.

Well, even a non-materialist like me thinks there are inherent lawlike connections between brain processes and consciousness. But for materialism to be true there has to be more than a lawlike connection -- witness the distinction between natural supervenience and logical supervenience.

>Oh, not to jump around but back to Fred, does it ruin everything if qualia
>is unique to each individual? Could it be that what you feel is your
>relation to that thing (so it involves both parties to the interaction)?
>

>Also, I can't help but think that Mary would dream in color-- does this
>count as a qualia?

I think for the purposes of the issues re materialism, it doesn't matter whether individual's qualia are unique or not. It may well be that our qualia are at least "colored" by individual factors.

Re Mary's dreams, you're right that this could well happen in practice (similarly she could presumably experience color patches when rubbing her eyes, and she'd see reddish qualities in her skin, etc). We just have to abstract away from those things for the purposes of the thought-experiment (or make Mary cortically colorblind, which would presumably do the job).

Anthony writes:

>In section IV of *Epiphenomenal Qualia* Jackson discusses three arguments
>that might be presented to show that qualia might have some role in
>determining brain states. I suppose that his commitment to the
>impossibility of qualia having any causal efficacy stems from the rest of
>his argument. If Mary were to know everything about physics,
>neurophysiology and so forth, but had not had the experience of seeing
>red, and seeing red had some causal role in determining brain states,
>then, presumably, she would not have all of the physical knowledge there
>is to have. There are two points I am not clear about here:

Well, I think Jackson is assuming that physics is causally closed. It's an interesting question what to say if interactionist dualism is true, i.e. if qualia are nonphysical and play a causal role with respect to a non-causally-closed physics. Maybe here one could still make a case that Mary could know (at least in principle) the physical facts about a system; she just wouldn't know about the nonphysical causes of some of the physical events.

Anyway, if interactionist dualism is true, materialism is already false, so for the purposes of the argument one can take it as a concession to the materialist to rule that view out. Of course it then becomes a further question whether, upon rejecting materialism, one should accept epiphenomenalism or interactionism. (Actually, I think of the choice as a threeway disjunction between epiphenomenalism, interactionism, and "panprotopsychism".)

>First, I am a bit confused about the phenomenal/ psychological division.
>It seems to me that the phenomenal quality of our experiences does play a
>causal role in determining behavior. Isn't it the feeling of pleasure that
>one has when one sees a particularly beautiful sunset that leads one to
>try to have such experiences? I think this is probably a misunderstanding
>on my part, but it does seem that when I want to eat a pint of Ben and
>Jerry's, it is the phenomenal quality of the experience that I find
>particularly appealing. How does this work if we speak in purely

>psychological terms?

To say that phenomenal concepts aren't functional concepts (or even to say that phenomenal states aren't functional states) doesn't automatically imply that phenomenal states don't play a causal role. It merely says, in effect, that they aren't **defined** by their causal roles. A concept such as "poison" is arguably a functional concept: to be a poison is roughly to be the sort of thing that has certain sickening effects on certain biological systems. So there's nothing more to the idea of a poison than the idea of something which plays a certain functional role. (Once Mary knows that something is the sort of thing that plays the role, she knows it's poison; one can't conceive of something being the sort of thing that plays the role without it being poison; etc, etc.) That's how it is arguably for very many concepts. But (according to the epistemic arguments) not so for qualia. Maybe these play a causal role, but that causal role doesn't define them.

The point can be put somewhat controversially by saying that qualia are intrinsic properties rather than functional properties. Functional properties are defined by their causal role. Intrinsic properties are defined by their intrinsic nature. Of course intrinsic properties can still play a causal role. Indeed, presumably any causal role will ultimately be played by some intrinsic state or property. E.g. the "poison" role in any given case will be played by such-and-such a chemical structure. It's plausible that the chemical properties in question are intrinsic properties. (N.B. There is a case that these may be functional properties at a lower-level of analysis, but I leave that aside for now.)

>This also gave me a worry in regard to the zombie argument. We are asked
>to admit the logical possibility of a world in which there are beings
>physically and functionally identical to us, but without consciousness.
>But why would these beings look at sunsets or eat ice cream? I don't know
>how behavior such as we exhibit is to be explained in terms of only
>functional states.

Well, the first thing to note is that it's not obvious that the same strictures of "explanation" hold in logically possible worlds as hold in the actual world. E.g. there are presumably possible worlds where monkeys coincidentally type "Hamlet", where the world explodes for no reason at all, etc. But anyway: presuming that actual physics is causally closed, I suppose the zombie's behavior might be explained simply in terms of its physical states. When it eats ice creams, certain receptors are triggered which leads to a state which it "likes", i.e. which leads it to attempt to return to the state again. It's not obvious that there's a problem in principle with explaining the behavioral part in physical terms. Of course one may find it counterintuitive that qualia don't play a causal role in our actual behavior, in which case there may be at least some reason for the nonmaterialist to consider alternatives on which they play such a role (such as interactionism or panprotopsyism).

>Secondly, is it a fact that prior to having a particular kind of sensory
>experience we are incapable of having experiences of that sort? Could
>Mary, having read about the experience of seeing red and having examined
>various MRIs and what not of people as they experience red, devise some
>sort of device to stimulate her brain in such a way as to have the
>phenomenal experience of red?

It's very likely that she could. But I don't think this affects the central dialectic, which concerns what Mary can know merely by virtue

of knowing the physical facts and reasoning about them. Direct brain stimulation in the production of experience is obviously a further "empirical" sort of knowledge. Crucially, the fact that Mary can do this in no way suggests that the phenomenal facts are entailed a priori by the physical facts. At best this relies on some sort of contingent a posteriori connection between them, the sort that even a non-materialist might embrace. (E.g., if there are contingent psychophysical laws connecting brain states to experiences, this method ought to work.)

>Sorry to be rambling somewhat. A somewhat unrelated question... In the
>first part of chapter you talk about the form of dualism you are
>advancing. You suggest that phenomenal properties are basic and that there
>must be fundamental psychophysical laws relating these phenomenal
>properties to physical properties. Will these laws relate particular
>physical (brain) states to certain conscious experiences? Is the point of
>this that it is a contingent matter what psychophysical laws actually
>exist?

Well, I think there will be psychophysical laws relating brain states to experiences; e.g., it will be a law that when you have such-and-such a brain state, you have an experience of red. But those laws probably won't be fundamental. It's unlikely that universal laws of nature will make reference to brains, etc. So one would expect there to be some much simpler underlying principles (e.g. principles linking information and experience?), from which the specific high-level laws are consequences. I take it that their nature is very much an open question, though.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Feb 7 21:15:06 1999

Date: Sun, 7 Feb 1999 22:04:33 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Rachael J Parkinson <rachaelp@U.ARIZONA.EDU>

Subject: more on reductive explanation

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Hello All-

Like Anthony, I hope I am not holding discussion back here...

In 3.3 of TCM, Chalmers criticizes attempts at reductive explanations of consciousness. In his critique of Baars cognitive model he argues, "But there is no reductive explanation of *experience* to be found here. The question of why these processes should give rise to experience is simply not addressed."(112) I take it that this is because the explanation of lower-level facts does not (and cannot) explain the higher-level phenomena of consciousness.

I am curious as to how this is different from the problem of reductive explanation in the case of the football game or the cocktail party. It does not seem that in giving a complete description of the lower level facts of a football game, I am really explaining how the lower level facts give rise to a football game. What is necessary for a complete explanation is not only a description of the lower level facts but also a description of the higher level phenomena (in terms of plays, players, touchdowns, etc.)

>From this, we may propose that what is necessary to supplement a cognitive or neurobiological model is a complete description of the higher level phenomena in question, specifically, consciousness. I assume that a

description of this sort would rely heavily on psychology. It seems that combining a reliable cognitive or neurobiological model with a comprehensive psychology should be able to play the same sort of explanatory role as describing a football game in terms of both the lower and higher level facts.

It is tempting to conclude that this sort of explanation of consciousness, in terms of lower level and higher level facts is satisfactory. However, objections to explanations of this type are given by Chalmers in this chapter. To take just one, we can conceive of something physically identical to ourselves not having consciousness (a zombie). But can we conceive of something physically identical to a football game not being a football game? Although it is conceivable that such a thing not be *called* a football game, it is safe to say that the higher level phenomena of a football game will be in place once all of the lower level facts are fixed.

My question is one which arises from Chalmers objection to cognitive models. Chalmers says, "The question of why these processes should give rise to experience is simply not addressed."(112) I am curious as to why and how these processes *should* give rise to experience, but as I have not read all of TCM, I take it the answer will surface at some point in my reading.

-Rachael

From owner-modality@LISTSERV.ARIZONA.EDU Mon Feb 8 04:45:34 1999
Date: Mon, 8 Feb 1999 04:45:30 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: more on reductive explanation
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts in responses to Rachael's and Erik L's notes on reductive explanation:

Rachael writes:

>I am curious as to how this is different from the problem of reductive
>explanation in the case of the football game or the cocktail party. It
>does not seem that in giving a complete description of the lower level
>facts of a football game, I am really explaining how the lower level facts
>give rise to a football game. What is necessary for a complete
>explanation is not only a description of the lower level facts but also a
>description of the higher level phenomena (in terms of plays, players,
>touchdowns, etc.)

Well, giving the low-level facts alone will never be enough for a reductive explanation, since you have to connect them to the high-level story. That's the part which is played by explication of the high-level concepts (e.g. as functional concepts), and by showing how the low-level properties can entail the high-level properties.

For functional concepts such as "poison", "life", and the like, this is fairly straightforward. Things are more complex for football games and cocktail parties, mostly because a complete explanation of these things centrally involves various *intentional* properties, characterizing people's attitudes, beliefs, desires, etc. And it's not at all obvious how these intentional properties are to be reductively explained. But leaving aside these worries stemming from

the "mental" aspects of the games and parties, it seems pretty clear that the residual high-level phenomena in question (the movement of people and balls and drinks, the production of utterances, etc) can be accounted for in physical terms. As before, it may not be the most illuminating explanation of all time, but we'll remove any ontological mystery.

The intentional aspects pose special worries, but here of course we're back to the mind-body problem itself. Personally I am hopeful that the intentional can be explained in terms of the functional and the phenomenal, but I don't say this is completely obvious. (N.B. The book is officially neutral about whether intentionality can be functionally explained, but I was probably more sympathetic to the idea then than I am now. I now tend to think that phenomenology may be central to a full understanding of intentionality.)

>From this, we may propose that what is necessary to supplement a cognitive
>or neurobiological model is a complete description of the higher level
>phenomena in question, specifically, consciousness. I assume that a
>description of this sort would rely heavily on psychology. It seems that
>combining a reliable cognitive or neurobiological model with a
>comprehensive psychology should be able to play the same sort of
>explanatory role as describing a football game in terms of both the lower
>and higher level facts.

Well, it's not just a matter of giving the low-level facts and the high-level facts at the same time. One needs to have the right sort of relationship of entailment between them. And as you point out below, it's this epistemic relationship that seems to be missing in the case of consciousness.

>My question is one which arises from Chalmers objection to cognitive
>models. Chalmers says, "The question of why these processes should give
>rise to experience is simply not addressed."(112) I am curious as to why
>and how these processes *should* give rise to experience, but as I have
>not read all of TCM, I take it the answer will surface at some point in my
>reading.

I'm not certain what the question is here. If it's about why and how the processes give rise to experience on my view, it's because there are fundamental psychophysical laws linking the physical and the phenomenal. At some point it's like "why do apples fall? Because of the law of gravity". The apple and the earth stand in the right sort of relation for the law to apply; and the brain is presumably in the right sort of state for psychophysical laws to apply. As to why the law exists, at some point one has to take that as a brute fact, just as in physics. Of course one wants the residual "brute" part to be as simple as possible.

Actually, we won't be talking much about these issues in the rest of the seminar, as we'll be focusing more on modality than on mind, but I'm sure some of these things will come up.

Erik writes:

>"not very
>illuminating" is perplexing in it's own right. What makes something
>"illuminating"? That's a strange notion once you've conceded that the
>process referred to is nothing over and above physics. I guess what I'm
>asking is, what is the role--ontologically, functionally, explanatorily or

>whatever--of concepts in explanations and why should they be NECESSARY if
>a supervenience relation holds between the phenomenon and a set of
>entirely physical facts? For that matter, WHAT IS A CONCEPT in this sense
>that we keep using it?

Re "illuminating", presumably this is an epistemic notion, and (unlike
"a priori") not an idealized one, so relative to human psychology. So
A can a priori entail B, but the entailment may not be terribly
illuminating to us, because we're not in a position to easily grasp
the deep structure in what is entailed, or something like that, at
least not without a lot of further work (and that further work is what
one might call "high-level explanation"). Whereas perhaps a more
cognitively skilled race of beings might see these things much more
easily.

Re the role of concepts: If the physical facts entail the B facts a
priori, then one who merely possesses the B concepts ought to be able
to figure out the B facts from the A facts. If this person can do
this merely in virtue of concept possession (plus reasoning) it seems
clear that the concept has to be of the right sort to enable the
entailment. In particular, there ought to be some sort of analysis of
the concept's conditions of application so we can see why they are
satisfied automatically whenever the A facts are satisfied.

The main point is that a priori entailment is an epistemic relation,
and epistemic relations seem to be grounded in relations between
concepts. The question of how these concepts (in our mind) correspond
to properties (in the world) and what this means about the bridge
between epistemology and ontology is one we'll be thinking about.

What's a concept? That's a hard one. We've been more or less helping
ourselves to the notion at an intuitive level, requiring only that
there is a concept corresponding to (most) words, that concepts have
conditions of application, and that these conditions of application
when composed yield the conditions of application of corresponding
sentences. Or something like that. The underlying metaphysics of
concepts is far from clear (though maybe the 2-D framework helps a
little in clarifying their structure). But I hope even the intuitive
level, thinking of concepts in terms of their pattern of application
to possibilities, is helpful for our purposes.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Feb 8 05:30:48 1999
Date: Mon, 8 Feb 1999 05:30:45 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Conceivability and Knowledge arguments
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Sorry to the slow reply to Tim's very interesting message about
conceivability.

First, Tim suggests that the knowledge argument may be a
conceivability argument, since one has to conceive of Mary in her lab
coat, etc. Personally, I think this act of conceiving is pretty
inessential (certainly I make not the slightest effort to really
conceive of Mary herself when thinking about this). What's more
central is the claim that conceivability or no, it's extremely
plausible that someone with full physical knowledge of the brain is

not thereby in a position to know what it's like to see red, and will learn something new when they see red for the first time. I don't think one needs to engage in detailed acts of conceiving to find this plausible.

The zombie case, on the other hand, at least involves a token attempt at conceiving the details of the case. I try to conceive of someone functioning just like me, behaving like me, with mechanisms of perceptual discrimination, attention and integration, higher-order monitoring, control of behavior, verbal report, etc, but no consciousness. Of course one's attention to detail is highly imperfect here, but as before there may be reasons to believe that the details can't make the crucial difference.

>But in fact, I doubt that we do anything **like** this in either case.
>Conceivability does **not** seem to involve the imagination - at least
>that's what it seems like to me in **these** cases. It's **not** like trying
>to work out in your head whether you can fit a ball of diameter two inches
>inside a triangle of sides 3, 4 and 4 inches which has an outer perimeter
>of 1/2 an inch thick.

I'm not sure exactly why you say this. Certainly there are some senses in which conceivability needn't involve imagination -- e.g. if one acquaints imagination with visual or sensory imagery. But it seems plausible to me that conceivability involves **something like** imagination -- e.g. a sort of "conceptual imagination", involving considering a lot of details about a case, and seeing if they are coherent and compossible.

>So what is going on in these cases? Well, it seems to me that in both
>cases you have the following intuitions or you don't.

>
>(a) Necessarily, all physical properties are functional or structural
>properties, and phenomenal properties aren't functional/ structural
>properties; OR, the very similar intuition: (b) Nec., all physical facts
>(or, perhaps, propositions) are exclusively composed out of functional
>/structural concepts, and phenomenal concepts are not
>structural/functional concepts.

>
>
>(a) is at work in the zombie cases, (b) is at work in the knowledge cases.
>If there is a one-one correspondence between facts and properties as is
>commonly supposed in setting these cases up, then it is hard to see why
>one should have (a) and not (b), or vice-versa. (Caveat: Of course, some
>do reject the claim that if you've got two facts you've also got two
>properties, but then you can accept that Mary does learn something new
>without rejecting physicalism.)

Hmm, this is interesting. It may well be the case that a tacit acceptance of something like your principles (a) and (b) underlies our acceptance of the conceivability and knowledge intuitions. But I think it would at least be highly tacit. Plenty of people would be unsure whether to accept the abstract principle, but quite confident about the conceivability or knowledge intuition. And it's certainly the case that a good number of people accept the Mary intuition while being uncertain about the zombie intuition, at least in its full-blown form. So I think there's reason to believe that something is going on in our evaluation of the conceivability and knowledge intuition that is at least somewhat different from each other and somewhat different from our evaluation of the abstract principles, though it may be the case that the three are a priori connectable and that someone who

accepts one should accept all three.

Incidentally I don't quite accept your diagnosis of (a) going with zombie intuitions and (b) with knowledge intuitions. It seems to me that (b) is weaker than (a), as it's a claim about concepts, not properties. The type-B materialist who holds that there is an epistemic gap but not an ontological gap may well accept (b) but not (a), for example. But this type-B materialist may well still have the zombie intuition, i.e. will find zombies conceivable; they'll just deny that conceivability implies possibility. So that suggests that (b), not (a), may underlie the zombie intuition, if anything does.

(Of course, one may call the "zombie intuition" the claim that zombies are metaphysically possible; then accepting the intuition may well go along with accepting (a). But then the zombie intuition isn't the pure epistemic intuition. I prefer to reserve the term "intuition" for the epistemic part, as the residual part is a matter of straight philosophy, not really intuition.)

Another slight difference between zombie and knowledge intuitions. Take someone who holds the epistemic theory of vagueness, according to which there is a sharp line in the application of vague predicates, although we can't know where that line is. So e.g. someone 5'10" is either tall or not tall (abstracting away from issues about context), but we can't know which through any amount of reasoning. Such a theorist will hold that complete physical knowledge won't enable someone (e.g. Mary) to know whether the person is tall. But it's not clear that such a person will hold that a 5'10" tall person and a 5'10" short person are both conceivable in the way that me and my zombie twin are both conceivable. Maybe they are "negatively conceivable" in that we can't rule them out a priori, but they don't seem to be "positively conceivable" in the sense that we can clearly and distinctly conceive of both scenarios. Rather there seems to be just one scenario that we conceive of, such that we can interpret it in two different ways.

So it seems that zombie intuitions involve a slightly stronger claim than knowledge intuitions. The former requires "positive conceivability", the latter mere "negative conceivability" or absence of contradiction. So maybe that is more evidence that the two are slightly distinct. The distinction between these two sorts of conceivability is something we'll be discussing quite a bit in a few weeks.

>But imagination does seem to play some role in conceivability. Consider
>the following problems: (1) is logically it possible to be an agent
>without, being a perceiver (or vice-versa)? (2) Is it possible for one
>subject of experience to have two distinct (spatially separated) bodies at
>the same time, so that, e.g., one of my bodies might be in Tucson, and the
>other one in Texas at the same time, and they are simultaneously
>conscious? Well, people seem to have very different intuitions about both
>of these cases? Why? Are they: (1) actually imagining different scenarios?
>or (2) imagining the same sorts of scenarios, but putting different labels
>on them because of the differing structure of their concepts? I dunno, but
>I'm tempted towards (2).

I'm tempted to agree with you here. It seems that at least in (1), the scenario one conceives of is pretty clear; the question is then how to describe and interpret it. One positively conceives a being who can't see/hear/etc, but who seems to move around, etc, and one considers whether that being is an agent. So we have a core conceived scenario and an issue of how to describe it. A little like the

epistemic vagueness case, actually, in that there's just one positively conceivable scenario in the vicinity. Contrast the zombie case, where there seem to be two (at least to some people -- at least here the issue is about whether there is one or two, rather than about how to interpret the obvious one). Maybe something similar goes in (2).

>Perhaps conceivability and knowledge arguments are different types of arguments because they are different ways of pumping the same intuitions (i.e. the same concepts), but I'm not even sure that this is true.

I tend to think they are ways of pumping similar intuitions, though perhaps not exactly the same intuitions. They certainly play separable roles as argument and a rhetorical devices, though of course that's often the case with "intuition pumps". As often in philosophy, there can be two closely related ways to reach the same conclusion,

I agree, though, that the key underlying point in both cases concerns the gap between our concepts. And the central questions are then (a) is there an a priori gap between the concepts and (b) if so does it lead to a gap in properties. The type-A materialist denies the gap in concepts. The type-B materialist accepts the gap in concepts but denies that it leads to a gap in properties. This type-B materialist will typically lean on a posteriori necessities ("water" and "H2O" are arguably different concepts picking out the same properties) and on Kripke's discussion thereof. Which leads nicely into this week's topic.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Feb 9 06:59:06 1999
x-sender: agillies@pop.u.arizona.edu
Date: Tue, 9 Feb 1999 08:10:51 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: twins and such
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

All,

Here's a puzzle I have. My trouble is in imagining my zombie twin.

(1) I think it's pretty clear that for purposes of the materialism debate it won't do for me to imagine (say) your zombie twin. You might *be* a zombie for all I know. Since I don't have any immediate access to your subjective states, a case where you don't consciousness isn't any different (from my vantage point) than a case where you do. So I have to imagine my own zombie twin.

(2) OK, now how do I do that? I imagine zombie twin earth. The problem is that there are lots of (zombie) folks there. How do I pick out my zombie twin? Since it's crucial that I pick out *my* zombie twin, I figure that I tag him *Zombie Thony*, and use the name to identify him. Imagining my zombie twin comes down to evaluating the truth of

(*) Necessarily, Thony is conscious
Materialism is false if (*) is false according to the 1-intensions.

(3) The trouble with this is (as I think you say in TCM) consciousness is at the core of our epistemic universe--we can't help but think we're conscious. But then the 1-intension of 'Thony' (when thought by me) will

include 'is conscious'. And so (*) is true after all.

Apparently something has gone wrong: my zombie twin *seems* possible, but in fact isn't when I try to imagine him. Conceivability and possibility seem to come apart. But I don't think things are really that simple. My guess is that one might be tempted to say that what I need to imagine is not really *my twin*, just any old thing that is physically identical to me. I'm probably missing something in this would-be reply. I thought that what's at stake is the truth of modal statements like (*), and imagining something physically identical to me doesn't decide that.

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Tue Feb 9 09:40:17 1999
Date: Tue, 9 Feb 1999 10:43:54 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Re: twins and such
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Reply to Thony on twins and such.

(Thony knows my line on this, but I thought it might be useful to make our discussion public.)

I think the false move (if there is one here) is the first move. Why won't it do to imagine somebody else's zombie twin? As far as I can see, it will do. I take it that the argument from zombiehood is meant to show that you can (logically) have all the functional/physical properties that we have, without having conscious experience. The reason why the zombie arguments ask you to imagine your own zombie, is that they hope that in this way they *guarantee* that you've got the sort of physical/functional properties on which consciousness naturally supervenes (since one supposes that one is oneself consciousness).

I read Thony's worry re imaging his zombie twin in the following way: if consciousness is one of my essential properties, then I don't have a zombie twin - there is no world in which there I exist and am not conscious (or, perhaps, capable of having conscious experience). I think Thony reads Dave as endorsing this premise or something like it. Whether or not Dave does, it's not an unreasonable assumption.

Suppose that we grant this point. I take it that the zombie argument can make do with *qualitative* identity rather than *numerical* identity. Suppose there is an organism just like you, twin Thony. Twin Thony is physically and functionally identical to you. That is, for every physical property that you instantiates, he also instantiates, and vice-versa. But we can conceive that twin Thony is not conscious, and if conceivability= logical possibility, consciousness does not (logically) supervene on the physical.

As I say, I think that you can probably run the argument without getting into worries about trans-world identity and individuation (which is good!): what matters is that isolate the sort of physical facts that the reductive materialist insists that consciousness is logically supervenient (and then you go to work with the zombie intuitions. . Talking about *My* zombie twin is just used (perhaps wrongly) as a

convenient way to isolate those facts.

Tim

'm not sure that the first step of

On Tue, 9 Feb 1999, Anthony S Gillies wrote:

```
> All,
>
> Here's a puzzle I have.  My trouble is in imagining my zombie twin.
>
> (1) I think it's pretty clear that for purposes of the materialism
> debate it won't do for me to imagine (say) your zombie twin.  You might
> *be* a zombie for all I know.  Since I don't have any immediate access to
> your subjective states, a case where you don't consciousness isn't any
> different (from my vantage point) than a case where you do.  So I have to
> imagine my own zombie twin.
>
> (2) OK, now how do I do that?  I imagine zombie twin earth.  The problem
> is that there are lots of (zombie) folks there.  How do I pick out my
> zombie twin?  Since it's crucial that I pick out *my* zombie twin, I
> figure that I tag him *Zombie Thony*, and use the name to identify him.
> Imagining my zombie twin comes down to evaluating the truth of
> (*) Necessarily, Thony is conscious
> Materialism is false if (*) is false according to the 1-intensions.
>
> (3) The trouble with this is (as I think you say in TCM) consciousness
> is at the core of our epistemic universe--we can't help but think we're
> conscious.  But then the 1-intension of 'Thony' (when thought by me) will
> include 'is conscious'.  And so (*) is true after all.
>
> Apparently something has gone wrong: my zombie twin *seems* possible, but
> in fact isn't when I try to imagine him.  Conceivability and possibility
> seem to come apart.  But I don't think things are really that simple.  My
> guess is that one might be tempted to say that what I need to imagine is
> not really *my twin*, just any old thing that is physically identical to
> me.  I'm probably missing something in this would-be reply.  I thought
> that what's at stake is the truth of modal statements like (*), and
> imagining something physically identical to me doesn't decide that.
>
> Thony
>
>
>
>
> "Curious green ideas sleep furiously."
```

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Tue Feb 9 11:25:50 1999
Date: Tue, 9 Feb 1999 12:25:07 -0700
Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik J Larson <erikl@U.ARIZONA.EDU>

Subject: Re: twins and such

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Some further thoughts on Thony and Tim's comments re "twins and such"

I admit it is puzzling to imagine imagining a physical replica of yourself without conscious experience. The gut feeling is something like, What would it mean to imagine MYSELF without my consciousness? I suppose in so far as one is trying to imagine one's twin "from within", there is no way to get the physical-replica-without-consciousness conception off the ground. So I suppose the conception of a zombie-twin or a physical replica of oneself sans consciousness when you want to construct it from the first-person point of view--that is the very perspective that is supposed to be excluded. So I suppose the job is to determine whether this perspective--the "from within" perspective"--can be detached from the rest of the picture in imagining a replica of oneself. I can imagine myself without an arm--all else being equal. And so on for all sorts of physical alterations. So the problem is not one of simply subtracting physical constituents--we can do that. Personality is trickier. Can one imagine ONESELF without certain memories, or dispositions, etc? Now I suppose that's a question of personal identity, and I think it is more puzzling than the case at hand. That is because, imagining yourself without conscious experience is tantamount to imagining yourself qua physical THING, or, imagining yourself as someone else would see you--from the third-person so to speak. And what is so tricky about that? We do it all the time, or can do it all the time, as even a moments reflection should convince one of. I personally do not have any trouble imagining myself qua physical thing--looking exactly the same, moving the same, behaving and reacting the same--only without any conscious experience whatever. Now I can't imagine ME in any strong sense this way, since I think that part of my personal identity is my consciousness, or that my consciousness is a precondition of having a personal identity, perhaps. So that is not a possibility. But, I think it is also not at issue here. What is at issue is whether we might conceive of ourselves qua physical thing, identical in all physical (and behavioral) aspects but lacking conscious experience. Would this be you in any strong sense? No. But would it be you in precisely the sense required--having the physical properties en masse but without conscious experience.

So, replying to Thony's point directly, I think he is conflating questions of personal identity with questions of purely physical conceivability, where the latter is less troublesome.

Erik L.

On Tue, 9 Feb 1999, Timothy J Bayne wrote:

> Reply to Thony on twins and such.

>

> (Thony knows my line on this, but I thought it might be useful to make our
> discussion public.)

>

> I think the false move (if there is one here) is the first move. Why
> won't it do to imagine somebody else's zombie twin? As far as I can see,
> it will do. I take it that the argument from zombiehood is meant to show
> that you can (logically) have all the functional/physical properties that

> we have, without having conscious experience. The reason why the zombie
> arguments ask you to imagine your own zombie, is that they hope that in
> this way they *guarentee* that you've got the sort of physical/functional
> properties on which consciousness naturally supervenes (since one supposes
> that one is oneself consciousness).
>
> I read Thony's worry re imaging his zombie twin in the following way: if
> consciousness is one of my essential properties, then I don't have a
> zombie twin - there is no world in which there I exist and am not
> conscious (or, perhaps, capable of having conscious experience). I think
> Thony reads Dave is endorsing this premise or something like it. Whether
> or not Dave does, it's not an unreasonable assumption.
>
> Suppose that we grant this point. I take is that the zombie argument can
> make do with *qualitative* identity rather than *numerical* identity.
> Suppose there is an organism just like you, twin Thony. Twin Thony is
> physically and functionally identical to you. That is, for every physical
> property that you instantiates, he also instantiates, and vice-versa. But
> we can conceive that twin Thony is not conscious, and if conceivability=
> logical possibility, consciousness does not (logically) supervene on the
> physical.
>
> As I say, I think that you can probably run the argument without
> getting into worries about trans-world identity and individuation (which
> is good!): what matters is that isolate the sort of physical facts
> that the reductive materialist insists that consciousness is logically
> supervenient (and then you go to work with the zombie intuitions. .
> Talking about *My* zombie twin is just used (perhaps wrongly) as a
> convenient way to isolate those facts.
>
>
> Tim
>
>
> 'm not sure that the first step of
>
> On Tue, 9 Feb 1999, Anthony S Gillies wrote:
>
> > All,
> >
> > Here's a puzzle I have. My trouble is in imagining my zombie twin.
> >
> > (1) I think it's pretty clear that for purposes of the materialism
> > debate it won't do for me to imagine (say) your zombie twin. You might
> > *be* a zombie for all I know. Since I don't have any immediate access to
> > your subjective states, a case where you don't consciousness isn't any
> > different (from my vantage point) than a case where you do. So I have to
> > imagine my own zombie twin.
> >
> > (2) OK, now how do I do that? I imagine zombie twin earth. The problem
> > is that there are lots of (zombie) folks there. How do I pick out my
> > zombie twin? Since it's crucial that I pick out *my* zombie twin, I
> > figure that I tag him *Zombie Thony*, and use the name to identify him.
> > Imagining my zombie twin comes down to evaluating the truth of
> > (*) Necessarily, Thony is conscious
> > Materialism is false if (*) is false according to the 1-intensions.
> >
> > (3) The trouble with this is (as I think you say in TCM) consciousness
> > is at the core of our epistemic universe--we can't help but think we're
> > conscious. But then the 1-intension of 'Thony' (when thought by me) will
> > include 'is conscious'. And so (*) is true after all.

> >
> > Apparently something has gone wrong: my zombie twin *seems* possible, but
> > in fact isn't when I try to imagine him. Conceivability and possibility
> > seem to come apart. But I don't think things are really that simple. My
> > guess is that one might be tempted to say that what I need to imagine is
> > not really *my twin*, just any old thing that is physically identical to
> > me. I'm probably missing something in this would-be reply. I thought
> > that what's at stake is the truth of modal statements like (*), and
> > imagining something physically identical to me doesn't decide that.
> >
> > Thony
> >
> >
> >
> >
> > "Curious green ideas sleep furiously."
> >
>
> Timothy J. Bayne
> RM. 213 Social Science
> Department of Philosophy
> University of Arizona
> Tucson, AZ 85721
> USA
>
> Hm ph. (520) 298 1930
>

From owner-modality@LISTSERV.ARIZONA.EDU Tue Feb 9 23:53:44 1999
Date: Tue, 9 Feb 1999 23:53:40 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: twins and such
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I'm sympathetic with what Tim and Erik say in reply to Thony's question, and I don't have a whole lot to add to their reply on the central point. But there are a number of very interesting issues in the vicinity, some of which tie in with current and forthcoming issues on Kripke and 2-D semantics, so here are a few thoughts.

In the conceivability argument, I don't think too much turns on whether the conceived being is one's own zombie twin. It can just as well be the zombie twin of any conscious being. Of course in one's own case there is the advantage that one is certain the actual being in question is conscious. But on the plausible assumption that other beings are conscious, the argument will work just as well for them.

Thony suggests a problem: imagining someone else to be a zombie won't be any different from imagining them to be conscious, because both will seem the same from my point of view. I think I disagree with this. There seem to be two pretty distinct epistemic possibilities that I can entertain re that being: that they are conscious or that they are not. It's true that both would "look" the same to me, but this just means that we not should individuate epistemic possibilities according to the way they look, or the way they seem, or according to the evidence one would have. I take it that to suppose that conceivability or epistemic possibility is determined by appearances or evidence involves a sort of verificationism; and I take it that verificationism is usually taken to be implausible precisely because

there seem to be distinct epistemic possibilities that would yield exactly the same evidence. E.g., empirically equivalent theories in science, or possibilities involving entities that leave no trace, or indeed possibilities involving consciousness.

So I think doing the thought-experiment on others is OK. Still, I suppose one doesn't want to be **forced** to do the thought-experiment only for others, as if one did, then presumably someone could evade the argument by accepting solipsism (i.e. denying that any other beings are conscious), and the conclusion of the argument would be "materialism is false or solipsism is true". Solipsism isn't exactly attractive, but this would be messy, so it would be best if the argument works for one's own twin.

Thony raises the worry: if it has to be **me** that is a zombie, maybe this can't happen, because consciousness may be part of the primary intension of our concept of ourselves, or some such. Here I would endorse what Tim and Erik say in reply. The epistemic argument against materialism merely involves the claim that " $P \rightarrow Q$ " is not a priori, or that P -and-not- Q is conceivable, where P is the complete physical truth about the world and Q is a phenomenal truth. Nothing here requires that the zombie world contains me, or that the zombie twin be me; it merely has to be a physical duplicate of our world, and he merely has to be a physical duplicate of me. And here, Thony's worry doesn't arise.

Still, it's an interesting question: could my zombie twin be me? Speaking for myself, I don't usually think of my zombie twin as being me; I think of him perhaps as some sort of close relative. But could I be a zombie? Matters depend here on whether one is thinking of standard "subjunctive possibility" (secondary-intension possibility) or "epistemic possibility" (primary-intension possibility).

In the subjunctive case, we're considering the standard sort of possibility that Kripke and others discuss. Here, I can imagine someone arguing that it is essential to sentient human beings (like me) that they be at least potentially sentient, so that the hypothesis that I could be a zombie is like the hypothesis (discussed by Kripke) that I could have descended from a different sperm and egg, etc. If so, the 2-intension of "me" and "DC", etc, can only pick out potentially conscious beings, and it's (secondarily) impossible that I be a zombie; counterfactual possibilities concerning a zombie should not be described as possibilities concerning me. (I say "potentially conscious" as it's quite possible that I be temporarily unconscious, and perhaps that I died before becoming conscious.) I'm not sure whether this essentialist claim is right, but it has a certain possibility.

The epistemic possibility (primary intension) case is closer to what matters in the epistemic arguments, though, since we're dealing with a priori connections, and in such epistemic matters, primary intensions are central. Is it epistemically possible that I'm a zombie? By this, we're in effect asking (modulo a few small frills): is there a conceivable scenario satisfying the primary intension of "I am a zombie"?

(Note that this isn't quite the standard sense of "epistemic possibility", where P is epistemically possible if it's true for all I know. Obviously, I know I'm not a zombie, so it isn't epistemically possible in that sense. Rather, I'm invoking the Kripke-esque sense in which it's epistemically possible that water is XYZ, even though we know that it isn't, and so on. We could call this "broad epistemic

possibility" -- it is broadly epistemically possible that P when P isn't ruled out a priori. If you don't like calling this "epistemic possibility", just think of it as "1-possibility" instead, though I think it is epistemic in some deep sense. We'll come back to this.)

On my view, it is 1-possible that I'm a zombie. There is a centered world centered on a zombie, and the primary intension of my term "I" picks out the zombie in that world, and the primary intension of my term "consciousness" doesn't apply to that being. Thony may not like this, on the grounds that the primary intensions of our conceptions of ourselves include consciousness. But on my view, that isn't quite right: the primary intension of my "I" concept just picks out the person at the center of a centered world. Of course I very strongly believe (even know) that I am conscious, but this isn't part of the semantics of "I". Similarly (but more so) for concepts such as "DC", "Thony", etc. So in a centered world centered on an unconscious being, the primary intension of "I" picks out that being.

One can come at this slightly differently, by forgetting the 2-D details for a moment, and asking directly whether it is broadly epistemically possible that I am a zombie, i.e. whether it is a priori that I am not a zombie, i.e. whether it is a priori that I am conscious. On my view, this isn't a priori. This is something we know, and know directly and in some sense "non-empirically", but we don't know it through reason alone; we know it in virtue of having certain experiences. I can't know that I am conscious independently of experience, so it isn't a priori. So I can still consider the broad epistemic possibility that I am not conscious (even though it is a possibility easily ruled out by things that I know), and I can use my concepts to describe that epistemic possibility, etc. The broad epistemic possibility in question (in effect, the hypothesis that my world is centered on a zombie) is describable as an instance of the epistemic possibility that I am a zombie, an instance of the possibility that I am not conscious, etc. So this is a 1-possibility, and is a broad epistemic possibility, even though I know it doesn't obtain, so it isn't an epistemic possibility in the standard sense.

So, on my view, it probably isn't 2-possible that I'm a zombie, but it is 1-possible that I'm a zombie. None of this matters much for the argument against materialism (for which the semantics of "I" is mostly tangential), but it's independently interesting, especially for the illustration it gives of the way 2-D semantics and modality works. Obviously this hooks up nicely with the Kripke discussion, and some of it turns out to be important later. What I say above certainly isn't trivial or obvious or uncontestable, so feel free to fire away with questions and reactions.

--Dave.

P.S. A few of these issues come up very briefly around p. 133 of the book (middle paragraph), where I suggest that (in effect to avoid the technicalities above which arise from considering epistemic possibilities in which I'm not conscious), it's easiest to consider conceivable scenarios in which I'm at the center, conscious, and in which some other people are zombies. So I suppose my official version runs with third-person zombies (and so officially has the solipsism loophole). In principle, I think things work even with zombies at the center, though I don't talk about that much in the book. The above gets at the beginning of how this works, though there are some more subtleties I haven't mentioned yet about whether concepts are required at the center of centered worlds. That turns out to be important for

various purposes, too.

P.P.S. More on the Kripke messages shortly. It would be good to get something from everybody soon (I have Thony's and Tim's contributions posted a couple of weeks ago already). I also look forward to getting a detailed summary of Tuesday's discussion (the more detailed the better).

From owner-modality@LISTSERV.ARIZONA.EDU Mon Feb 8 21:40:06 1999
Date: Mon, 8 Feb 1999 22:28:53 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Angela J Burnette <aburnett@U.ARIZONA.EDU>
Subject: Re: Kripke's arguments for a priori contingencies
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Regarding Kripke's arguments for a priori contingencies and a posteriori necessities, I am assuming that the basis for the arguments is the fact that a prioricity is an epistemic notion while necessity is a metaphysical notion. Given this, there is the possibility that necessary truths be known only a posteriori and contingent truths (if true) can be known a priori. Identity statements such as Hesperus is Phosphorus are supposed to be examples of the former because, while the identity (which is between rigid designators?) is necessarily true, we could only find this out through empirical means. This sounds plausible, but I have a problem with Kripke's example of an a priori contingent truth...

He argues that the length of S at time 0 is used to fix the referent of the rigid designator 'one metre' by way of definition, which is not to say that the MEANING of 'one metre' is 'the length of S at time 0'; so, even though we can know that 'the length of S at time 0 is one metre' a priori (because that's how we defined it?) this is a contingent truth because the length of S at time 0 could have been different (if it were heated or whatever). Now, 'one metre' is a rigid designator, while 'the length of S at time 0' is not; so the meaning of 'one meter' is not the same as the meaning of 'the length of S at time 0'. But, given this, what exactly is the meaning of 'one metre' if, as Kripke stipulates, the length of S at time 0 could be the only standard being used? In other words, given that there is no standard for 'one metre' besides 'the length of S at time 0' how can 'one metre' be a rigid designator? How could someone reapply the term without reference to the description which is only supposed to fix its reference, not provide its meaning? I don't see why this isn't a problem. If someone can tell me where I've gone wrong in describing Kripke's account, such that I'm asking this question, I'd appreciate it.

angela

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 00:47:02 1999
Date: Wed, 10 Feb 1999 00:46:58 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Kripke's arguments for a priori contingencies
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Angela writes:

>Regarding Kripke's arguments for a priori contingencies and a posteriori
>necessities, I am assuming that the basis for the arguments is the fact
>that a prioricity is an epistemic notion while necessity is a metaphysical
>notion. Given this, there is the possibility that necessary truths
>be known only a posteriori and contingent truths (if true) can be known a
>priori. Identity statements such as Hesperus is Phosphorus are supposed
>to be examples of the former because, while the identity (which is between
>rigid designators?) is necessarily true, we could only find this out
>through empirical means. This sounds plausible, but I have a problem with
>Kripke's example of an a priori contingent truth...
>

>He argues that the length of S at time 0 is used to fix the referent of
>the rigid designator 'one metre' by way of definition, which is not to say
>that the MEANING of 'one metre' is 'the length of S at time 0'; so, even
>though we can know that 'the length of S at time 0 is one metre' a priori
>(because that's how we defined it?) this is a contingent truth because the
>length of S at time 0 could have been different (if it were heated or
>whatever). Now, 'one metre' is a rigid designator, while 'the length of S
>at time 0' is not; so the meaning of 'one meter' is not the same as the
>meaning of 'the length of S at time 0'. But, given this, what exactly is
>the meaning of 'one metre' if, as Kripke stipulates, the length of S at
>time 0 could be the only standard being used? In other words, given that
>there is no standard for 'one metre' besides 'the length of S at time 0'
>how can 'one metre' be a rigid designator? How could someone reapply the
>term without reference to the description which is only supposed to fix
>its reference, not provide its meaning? I don't see why this isn't a
>problem. If someone can tell me where I've gone wrong in describing
>Kripke's account, such that I'm asking this question, I'd appreciate it.

The crucial part of Kripke's account is that when a term is a rigid designator, once one has fixed its reference to a particular object (e.g. "Hesperus" refers to the planet Venus) or to a particular property (e.g. "hot" refers to a certain energy property) or to a certain length (e.g. "one meter" refers to length L, about 39.37 inches), one can then consider modal questions about what the term picks out in alternative possible worlds just by considering what is "that object" or "that property" or "that length" in other possible worlds, and by forgetting about the factors that were involved in fixing reference. So "Hesperus" picks out Venus in all worlds, irrespective of whether Venus is the evening star there. Similarly, one meter picks out length L in all worlds, irrespective whether the stick in Paris has that length there. The reference-fixer serves as a sort of bootstrap, or as a ladder we can kick away once it has done the job of giving the reference.

The reference of "one meter" is fixed by looking at what is the length of stick S, but once we have accepted that this is a given length L, then we can consider what is length L in other worlds quite independently of what is the length of S there. And according to Kripke, the modal application conditions of "one meter" go with the former (length L), not with the latter (the length of S). As Kripke notes, it seems that we can quite coherently consider counterfactual possibilities in which the length of stick S was less than one meter (just say it broke, or that it had never grown that long in the first place); so the application of "one meter" to these possibilities can't merely involve the length of stick S.

(Incidentally, if you accept that it's necessary that Hesperus is Venus, then presumably you don't think it's necessary that Hesperus is the evening star (as it certainly isn't necessary that Venus is the evening star). But arguably it's a priori that Hesperus is the evening star (by analogy with the meter case). So you'll then be committed to something like a contingent a priori statement there, i.e. "Hesperus is the evening star".)

(One difference between the meter case and the Hesperus case is that one might argue that there are deeper problems with reidentifying lengths across worlds than with reidentifying planets across worlds. What counts as length L in a given world? In reply, I think Kripke would make his suggestion here that we can "stipulate" lengths in our possibilities (e.g. considering the possibility that something is two meters long) rather than have to "find" the length through a distant

telescope. But there are certainly some issues here.)

What's the "meaning" of one meter? Kripke manages to stay away from that sort of question for the most part. Many people take the moral of Kripke's discussion to be that the "meaning" of the term is distinct from what "fixes reference" to the term. The reference of "one meter" is fixed as "whatever is the length of stick S"; but the "meaning" of the term involves its condition of application to possible worlds, which is different. The term "one meter" picks out length L in all worlds, so the "meaning" of the term involves that every length. Analogously, one might think of the reference-fixation of "Hesperus" as going via whatever is the evening star, but the "meaning" as involving the planet Venus. This is the sort of thing a "direct reference" theorist might say, for example, and there are a couple of passages in Kripke that suggest sympathy (e.g. where he distinguishes "fixing reference" from "giving the meaning").

For my part, I think both reference-fixing conditions and counterfactual application-conditions are in some sense part of the "meaning" of a term. I think of the former as the epistemic application-conditions (primary intension) and the latter as the subjunctive application-conditions (secondary intension). The former can be thought of as the a priori part of meaning; the latter as the (often) a posteriori part. In the 2-D framework, we'll say (roughly) that the primary intension of "one meter" picks out whatever is the length of a certain stick in all centered worlds, and that the secondary intension picks out a specific length L in all worlds. It doesn't matter much which of these we call the "meaning", but I think one can make a case that both are crucial.

If one likes, one can think of the a priori conceptual analysis of "one meter" as something like "the actual length of stick S" or "dthat(the length of stick S)". Then we can make sense of most of the phenomena. E.g., we can make sense of Kripke's counterfactual possibility that stick S could be less than one meter by noting that stick S could have been shorter than its actual length (i.e., than its length in the actual world). i.e., there are worlds in which the secondary intension of "one meter" and the secondary intension of "the length of stick S" come apart. But it isn't epistemically possible that stick S is less than one meter, as the primary intensions of "one meter" and "the length of stick S" are the same.

--Dave.

From agillies@pop.u.arizona.edu Wed Jan 27 09:22:44 1999
Subject: initial discussion--names and intensions
Date: Wed, 27 Jan 1999 10:35:36 -0700
x-sender: agillies@pop.u.arizona.edu
From: Anthony S Gillies <agillies@u.arizona.edu>
Status: RO

Hi all,

I was wondering about applying the two-dimensional framework to the arguments that Kripke gives in *N & N*, particularly those against the cluster concept view of names in the first lecture. Does it seem fair to say that the cluster concept view (a la Searle) is a theory about the primary intensions of names, and Kripke's examples point to the secondary intensions of names?

For example: the primary intension of 'Aristotle' can be cashed out as the guy who has some chunk of the properties: pupil of Plato, teacher of

Alexander the Great, and so on. This function won't pick out the same guy in all possible worlds, and so "Aristotle was a pupil of Plato" will be contingently true (according to the 1-intensions).

Kripke's examples, though, go like this: OK, imagine a counterfactual world where Aristotle was not a pupil of Plato (or any old property we think picks out Aristotle in this world). It is still a counterfactual situation about *Aristotle*. So, the name 'Aristotle' picks out the guy Aristotle in all possible worlds (according to the 2-intensions). So names are rigid designators.

What this suggests, I think, is that assuming the two-dimensional picture of reference, Kripke's arguments about how names work is missing a premise: namely, that in natural language we mostly care about the 2-intension of names. So: (1) does this seem like an accurate description of the landscape when we apply two-dimensional semantics to $_N$ & $N_?$; and (2) do we ever care about the primary intensions of names?

--Thony

P.S. I've BCC'ed this email to everyone in the class, that way people don't need to scroll through so many address headers at the beginning of every email.

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 01:48:56 1999
Date: Wed, 10 Feb 1999 01:48:53 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Names and intensions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Here's a reply to Thony's message from two weeks ago re Kripke.

>From agillies@pop.u.arizona.edu Wed Jan 27 09:22:44 1999
>Subject: initial discussion--names and intensions
>Date: Wed, 27 Jan 1999 10:35:36 -0700
>From: Anthony S Gillies <agillies@u.arizona.edu>
>
>Hi all,
>
>I was wondering about applying the two-dimensional framework to the
>arguments that Kripke gives in $_N$ & $N_?$, particularly those against the
>cluster concept view of names in the first lecture. Does it seem fair to
>say that the cluster concept view (a la Searle) is a theory about the
>primary intensions of names, and Kripke's examples point to the secondary
>intensions of names?
>
>For example: the primary intension of 'Aristotle' can be cashed out as
>the guy who has some chunk of the properties: pupil of Plato, teacher of
>Alexander the Great, and so on. This function won't pick out the same
>guy in all possible worlds, and so "Aristotle was a pupil of Plato" will
>be contingently true (according to the 1-intensions).
>
>Kripke's examples, though, go like this: OK, imagine a counterfactual
>world where Aristotle was not a pupil of Plato (or any old property we

>think picks out Aristotle in this world). It is still a counterfactual
>situation about *Aristotle*. So, the name 'Aristotle' picks out the guy
>Aristotle in all possible worlds (according to the 2-intensions). So
>names are rigid designators.
>
>What this suggests, I think, is that assuming the two-dimensional picture
>of reference, Kripke's arguments about how names work is missing a
>premise: namely, that in natural language we mostly care about the
>2-intension of names. So: (1) does this seem like an accurate
>description of the landscape when we apply two-dimensional semantics to
>_N & N_?; and (2) do we ever care about the primary intensions of names?

First a bit of background, which may already be clear to some people,
but which is very important and makes some of these issues clearer.

There are actually two very distinct strands in Kripke's argument
against the traditional view of names.

The first, centered on Lecture 1, is what we can call the "modal
argument". This makes the point that names are rigid designators, so
that their counterfactual conditions of application are different from
those of the various descriptions in the vicinity. E.g., for any
description D, there's a counterfactual possibility that Aristotle
wasn't D.

The second, centered on Lecture 2, is what we can call the "epistemic
argument". Here, he argues that it's not even the case that the
reference of names is fixed via description. This is where he brings
in the cases of Feynman, Godel, Jonah, etc, and argues that for any
description D in the vicinity, it is not a priori that Feynman (etc.)
satisfies D. That's because it's always epistemically possible that
Feynman didn't satisfy D; we could imagine finding that out. So not
only can't we be descriptivists about modal application conditions, we
can't even be descriptivists about reference-fixing.

To see that the two points are quite distinct, note that the modal
argument alone is compatible with the idea that names are "descriptive
names": their reference is fixed via some description, but then they
pick out the referent in question in all possible worlds. Some names,
are arguably like this: e.g. "Jack the Ripper" arguably has its
reference fixed to whoever committed the murders, in effect via
description, but then picks out that very person (person X) in all
possible worlds. Note that for a name such as "Jack the Ripper", one
could run the modal argument, but not the epistemic argument: it's not
necessary that Jack the Ripper committed the murders (there are worlds
where someone stopped him), but it's a priori that Jack the Ripper (if
he exists) committed those murders. But Kripke argues that most names
are not like this. This is why (in Chapter 2) he argues that the
descriptive theory of reference has to be supplanted by a causal
theory (or picture) of reference.

The modal point, in effect, is a point about secondary intensions. It
concerns the application conditions of our concepts to counterfactual
scenarios, given that the actual world is fixed. And we can express
Kripke's insight by saying that the secondary intensions of names pick
out the same object in all worlds. E.g. the secondary intension of
"Jack the Ripper" picks out the same person in all worlds,
irrespective of whether that person committed any murders there.

The epistemic point, in effect, is a point about primary intensions.
It concerns the application of our concepts to epistemic
possibilities, and concerns matters of reference-fixation. Even

though the secondary intension of "Jack the Ripper" goes with a specific person, not the description, the primary intension goes with the description: the primary intension of "Jack the Ripper" picks out (roughly) whoever committed the murders in question in a given centered world. But Kripke's point is that most names are not like this. We can express it by saying that the primary intensions of most names are not captured by any description. Rather, the primary intension involves picking out whatever is at the other end of a long causal chain.

Kripke's epistemic argument is somewhat more controversial than his modal argument, and one might argue that there are certain resources available to a description theorist to resist his conclusion. E.g., one might invoke descriptions such as "the entity at the other end of an appropriate causal chain" (causal descriptivism), or "the person called such-and-such around here" (metalinguistic descriptivism). And arguably the 2-D framework can give support to a certain sort of attenuated descriptivism. But that's a complex issue (feel free to raise issues in the area). Kripke's epistemic point re names is widely but not universally accepted in contemporary philosophy; I think it's fair to say that the modal point is fairly close to being universally accepted these days (or as close to universal acceptance as anything gets).

Now to Thony's points (some of which are in effect addressed already). Re the cluster theory: pre-Kripke, the modal and epistemic issues weren't carefully distinguished. It may be that in effect, Searle intended this as a theory of both, and in effect was concerned with both primary and secondary intensions. But I think it's arguable that it was most centrally intended as a theory of what fixes actual reference, and so was implicitly concerned with primary intensions. Certainly, the modal argument seems fairly decisive against it as a theory of secondary intensions; one might think it has more promise as a theory of primary intensions. But Kripke in lecture 2 is concerned to argue against it even as a theory of primary intensions (via the epistemic arguments). Someone like Searle might resist by accepting the modal point but not the epistemic point, and arguing that one can still give some sort of modified cluster theory at least as a theory of reference-fixation, if not as a theory of counterfactual application. That sort of debate is to some extent still going today.

But anyway, I don't think it's quite right to say that Kripke isn't concerned with the primary intensions of names, given all the discussion in lecture 2. And I think that one could argue that the "missing premise" of his argument against description theories is supplied by the argument in lecture 2. It's true, though, that the modal discussion in lecture 1, which is arguably the central point, is largely concerned with secondary intensions.

I think you're right that in some sense, Kripke thinks we are most concerned with secondary intensions of names rather than primary intensions. As noted before, he seems to implicitly see SIs as more closely tied to "meaning", whereas the "reference fixation" involved in PIs isn't really a matter of meaning. Of course matters are complicated by the fact that Kripke doesn't put things in terms of the 2-D framework at all (I'm told he isn't all that keen on it). Still, it's easy to translate much of his discussion into those terms.

One further reason for his preference for SIs over PIs may be what we might call the "semantic argument". Arguably, different users of the same name can have quite different primary intensions for it, though

the same secondary intensions. E.g., maybe my PI for "Bill Clinton" is different from Hillary's PI for "Bill Clinton", though we have the same SI. And more generally, maybe the PI of a name can vary throughout a community, whereas the SI is constant. If so, and if one adds the premise that the meaning of a term should be something that is shared between users of that term, one can see the case for thinking of the SI, not the PI, as part of a term's "meaning".

I think this is one of the reasons why contemporary philosophers of language tend to focus more on names' SIs, not PIs. (Another reason may be that some are skeptical of the very idea of a PI, or haven't thought much about it.) For my part, I think that even if one accepts that the PI of a term may not be the same for all users, there's still a sense in which it is part of the meaning of a term at least for a specific user (and arguably, it corresponds much more closely than the SI to what we can think of as the "cognitive content" of the term for the user). Or one can distinguish between "sentence meaning" and "utterance meaning" (as Marga Reimer does in some of her work) and argue that PIs are at least part of the latter, perhaps. But again, what truly gets to count as "meaning" is to some extent a terminological issue.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Jan 31 23:18:29 1999
Date: Mon, 1 Feb 1999 00:07:19 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.Arizona.EDU>
Subject: Comment on Thony; thoughts on primary intensions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts. . .first, a comment on Thony's message, then a thought or two on primary intensions. It's pretty long - sorry.

(1) Some comments about Thony's point re the 2 level framework and Kripke's ruminations on proper names. (a) First, an exegetical point. I find it very difficult to nail down the distinction between Searle and Kripke. When you read them very closely (i.e. footnotes and all), you see that Kripke allows for description to play a role in fixing reference, and Searle allows for causal chains to play a role in transmitting reference, and it becomes harder to see where they really disagree. (b) Assuming that a 2 level approach to proper names might work, Thony asks whether we ever care about the primary intension of proper names (i.e. the Searlian descriptive part). Note that Chalmers seems to think that we care about primary intentions *more* than secondary intensions - they are 'cognitive content' of thought (p. 65). I'm tempted to say that our use of a proper name (drawing inferences from it, etc.) involves its primary intension - and thus, at the very least, we do care about the primary intension a lot. But there may be another sense in which we also care about secondary intensions (i.e. the referent itself). (c) A disanalogy between natural kind terms and proper names is that the primary intension of the former is presumably shared by all members of the linguistic community who have the term (even if only by relations of deference, e.g. "elm"), while the descriptive content that I associate with, say, "Bill Clinton" may not be that which others, say ex-interns, associate with that same proper name. I'm not sure whether or not this difference is a difference that makes a difference.

(2) (a) Chalmers says that primary intensions may not always be easy to

uncover, but they are in principle open to reflection (to the degree that they are determinate). Here's a couple of cases that people have wildly different intuitions about and reflection really doesn't seem to help all that much. (i) I could have been a builder; (ii) I could have been born a century earlier than I was. (iii) I could have been a poached egg. (That last one is meant to be read literally - and it is in fact false in case you are wondering). (iv) I could have been the number 2. Some think that only (i) is true. Some think that (i) and (ii) alone are true. Some think that only (i), (ii) and (iii) are true. Perhaps you think that all four are true. (In which case you are a poached egg.) Is this just a particularly difficult case for the notion of the in principle transparency of primary intension, or does it suggest deeper and more pervasive problems with it?

This example also raises another question. Some want to say that it shows that some of my properties are essential to me (perhaps, say, being human), while other properties are not. How do we cash out the difference between these two types of properties in the 2D framework? Perhaps it goes like this. For any possible world P and any property x, if I have x essentially, then it is logically true that I have x in every world in which I exist, and if I have y only contingently, then it is logically true that there are some world in which I exist and have y, and other worlds in which I exist and don't have y, and which worlds are which depends on the natural laws and boundary conditions therein. In this way, perhaps, we can cash out the difference between essential and non-essential properties, without bringing in metaphysical necessity.

(b) Chalmers responds to the Quinean point about revisability by saying that his arguments are not affected, because they turn on supervenience conditionals (p. 55). I'm not sure why the fact that the propositions in question are supervenient conditionals should ipso facto bestow on them a certain kind of immunity to revision. First, note that our judgments about laws of nature (i.e. conditionals) are revisable. Second, note that we can distinguish two types of supervenience conditionals. First, there are those conditionals that state that B facts merely are supervenient on some other range of facts, without specifying which range of facts forms the supervenience base. (Actually, this is not really a supervenience conditional, but rather the claim that there is such a conditional.) Second, there are those conditionals which go on to say that B facts are supervenient on A facts. Now, presumably we can be wrong about the latter type of supervenience conditional. It is harder to see how we could be wrong in thinking that a certain type of fact is supervenient. We might go wrong in thinking that, say, aesthetic facts supervene on F facts, but could we err in thinking that aesthetic facts are supervenient in the first place? Chalmers, I take it, claims that we can't go wrong in such have an argument here, but it seems to me that we could mistakenly assume that a certain type of fact is supervenient although it is fact not, and vice-versa (its the vice-versa that DC needs, I think).

(c) I think that Chalmers wants to tie in the claim about supervenience conditionals being unrevisable, with the claim that we have transparent access to the primary intensions of our concepts, but I'm not sure that I see the connection. It seems to me that primary intensions are something like application conditions - this is what "watery stuff" seems to come to. But need application conditions involve a conception of whether or not a property is supervenient or not? I don't see it. One's general conception of what is or is not supervenient, and what may be supervenient on what, seems to be a fairly sophisticated thing. But perhaps that's okay. Perhaps all Chalmers needs to say is that if you have the (note: not "my") concept of consciousness, then you have the ability to reflect on it, and to see that it could not be logically supervenient on any other

types of properties. So: one can have a perfectly good concept of consciousness without having grasp of the supervenience conditionals that is contained within it and are putatively immune to Quine's worries. Is this the connection?

Sorry to be so long-winded.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 02:37:53 1999
Date: Wed, 10 Feb 1999 02:37:47 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Comment on Thony; thoughts on primary intensions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on Tim's old message.

>Date: Mon, 1 Feb 1999 00:07:19 -0700
>From: Timothy J Bayne <bayne@U.Arizona.EDU>
>Subject: Comment on Thony; thoughts on primary intensions
>To: MODALITY@LISTSERV.ARIZONA.EDU
>
>Some thoughts. . .first, a comment on Thony's message, then a thought or
>two on primary intensions. It's pretty long - sorry.
>
>(1) Some comments about Thony's point re the 2 level framework and
>Kripke's ruminations on proper names. (a) First, an exegetical point. I
>find it very difficult to nail down the distinction between Searle and
>Kripke. When you read them very closely (i.e. footnotes and all), you see
>that Kripke allows for description to play a role in fixing reference, and
>Searle allows for causal chains to play a role in transmitting reference,
>and it becomes harder to see where they really disagree. (b) Assuming that
>a 2 level approach to proper names might work, Thony asks whether we ever
>care about the primary intension of proper names (i.e. the Searlian
>descriptive part). Note that Chalmers seems to think that we care about
>primary intentions *more* than secondary intensions - they are 'cognitive
>content' of thought (p. 65). I'm tempted to say that our use of a proper
>name (drawing inferences from it, etc.) involves its primary intension -
>and thus, at the very least, we do care about the primary intension a lot.
>But there may be another sense in which we also care about secondary
>intensions (i.e. the referent itself). (c) A disanalogy between natural
>kind terms and proper names is that the primary intension of the former is
>presumably shared by all members of the linguistic community who have the
>term (even if only by relations of deference, e.g. "elm"), while the
>descriptive content that I associate with, say, "Bill Clinton" may not be
>that which others, say ex-interns, associate with that same proper name.
>I'm not sure whether or not this difference is a difference that makes a
>difference.

Re (a), I think you're right that Searle's and Kripke's views of reference-fixing (i.e. primary intensions) are closer than they may seem. I think the crucial difference at that while Kripke allows descriptions a role, you can't do everything with descriptions, whereas Searle thinks you can. Of course, as you note, some of Searle's descriptions may involve "the thing at the other end of a causal chain", or something like that. This leads to what is sometimes called "causal descriptivism" (a sort of hybrid); but I think it's a view that is more congenial to the descriptivist in some ways (after all, why should they object to having some descriptions involve causation). Kripke's response (e.g. around p. 162, and footnote 38) is to suggest that this is a pretty trivial sort of descriptivism (incorporating the whole causal theory as a part!), but the issue isn't entirely clear. Certainly someone like David Lewis is happy to take this on board as enough descriptivism for the descriptivist's central purposes (though it may be somewhat unlike a traditional descriptivism).

Re (b), yes, my own view is that primary intensions are in many ways more important than secondary intensions. I think they are more central to thought, to rational inference, and to capturing the way that language describes the world. On my view secondary intensions are mostly crucial to making sense of counterfactual thought and language; that's an important role but somewhat in the background. (Just why contemporary philosophy has then focused so much on SIs becomes something of a question; we'll talk about that a bit in the upcoming discussion of "the tyranny of the subjunctive"). Of course reference is important, but I think it is important qua reference, not qua its role in constituting an SI (to see this, note that reference is just as important for descriptions, like "the inventor of the zip", even though in these cases the SI doesn't involve the actual referent). But all this is just my view and of course is arguable.

Re (c), I discussed this in the reply to Thony. I think it is plausible that PIs of names, but not SIs, vary in a community. It's actually not entirely obvious whether or not the same goes for natural kind terms. I think someone might argue that different people could have different PIs for "water" or "gold" or whatever. For example, maybe there could be an isolated chemist who only associated "water" with the "H2O" description, and doesn't know about its appearance. His PI would be different from ours, but arguably he would still be using the term "water". Someone might suggest in reply that he doesn't really have our word, as the "watery" mode of presentation is somehow essential to the word; the resolution of that debate isn't entirely clear. But in any case it seems that at least mild variations in PIs may be possible even in the natural kind case.

>(2) (a) Chalmers says that primary intensions may not always be easy to
>uncover, but they are in principle open to reflection (to the degree that
>they are determinate). Here's a couple of cases that people have wildly
>different intuitions about and reflection really doesn't seem to help all
>that much. (i) I could have been a builder; (ii) I could have been born a
>century earlier than I was. (iii) I could have been a poached egg. (That
>last one is meant to be read literally - and it is in fact false in case
>you are wondering). (iv) I could have been the number 2. Some think that
>only (i) is true. Some think that (i) and (ii) alone are true. Some think
>that only (i), (ii) and (iii) are true. Perhaps you think that all four
>are true. (In which case you are a poached egg.) Is this just a
>particularly difficult case for the notion of the in principle
>transparency of primary intension, or does it suggest deeper and more

>pervasive problems with it?

Hmm, I think you're concerned here not with primary intensions but with secondary intensions. You're considering subjunctive counterfactual possibilities about what could have been; i.e., the sort of modal possibilities Kripke is concerned with in N&N (especially lecture 1). That is, I take it that people who are concerned with these questions are worrying about the same sort of modality as in Kripke's questions about whether I could have descended from a different sperm and egg (given that I descended from these ones). These are matters of secondary intension.

Personally, I don't have really strong intuitions about what I (DC) could and couldn't have been in various counterfactual possibilities. Maybe I could have been born earlier; probably not a poached egg or the number 2, but I'm not certain. This goes along with the idea that the secondary intensions of many of our concepts are not all that determinate; on my view, it is often almost a terminological issue which way to go. (There are even times when I'm tempted by the contra-Kripkean thought that water could have been XYZ.) Of course other claim to have much more determinate intuitions about these things, and even when they don't, they think that at least the SI is fairly determinate.

(Of course it's not required that we have a priori access to SIs. But one does hope that one has a priori access to SIs conditional on knowing the nonmodal facts (e.g. about actual reference). So insofar as SIs still seem a bit loose even given the nonmodal facts, it ought to be the case that either (a) one hasn't analyzed the case well enough, or (b) the SI is in fact indeterminate, or (c) disagreeing people actually have slightly different concepts.)

As to primary intensions, I think these are generally a lot more determinate and accessible, though of course even these can be fuzzy around the edges. When one is considering not subjunctive possibilities but epistemic possibilities about the way the world is, I think our terms often seem to have a pretty determinate application. There may be variation still in borderline cases and in very way-out cases, but there's at least a fairly determinate core; and much of the variation in wild cases can be dismissed as somewhat terminological and so not really important metaphysically. Again, I'll want to make the claim of a priori access, and argue that disagreement involves either (a) inadequate reflection, (b) indeterminacy in PI, or (c) terminological differences.

>This example also raises another question. Some want to say that it shows
>that some of my properties are essential to me (perhaps, say, being
>human), while other properties are not. How do we cash out the difference
>between these two types of properties in the 2D framework? Perhaps it goes
>like this. For any possible world P and any property x, if I have x
>essentially, then it is logically true that I have x in every world in
>which I exist, and if I have y only contingently, then it is logically
>>true that there are some world in which I exist and have y, and other
>worlds in which I exist and don't have y, and which worlds are which
>depends on the natural laws and boundary conditions therein. In this way,
>perhaps, we can cash out the difference between essential and
>non-essential properties, without bringing in metaphysical necessity.

I think the difference between essential and inessential properties will show up in the SI of a term. We'll say that Fred has property P essentially if "Fred is P" is necessary, i.e. 2-necessary (necessary in the subjunctive sense that Kripke considers). This will hold if

the SI of "Fred" picks out a being with P in all worlds. Fred has P inessentially if the SI of "Fred" picks out a being without P in some worlds.

In a certain sense, this is cashing out the difference between essential and inessential properties without bringing in any really heavy notion of metaphysical necessity. One still needs the basic space of worlds, but these can just be the conceivable worlds, more or less. Once one has worlds plus SIs, one has enough to understand the distinction. To get SIs, one needs (a) a good enough a priori grasp of the concept in question, and (b) nonmodal facts about the actual world. (E.g., say water is essentially H₂O. Then if one has a full a priori competence with the term "water" and knows the nonmodal facts about the composition of actual water being H₂O, one can know that the SI picks out H₂O in all worlds. Etc.)

Indeed, it seems pretty clear that Kripke's methodology for thinking about what's essential is more or less like this. The only empirical knowledge he needs are some nonmodal facts, e.g. that he's descended from sperm S and egg E, or that water is made of H₂O, and so on. The rest seems to follow by a priori analysis: arguably, by the method of conceivability plus a priori analysis of the way our concepts apply to worlds.

If this works, we can cash out essentialism relying only on (a) an a priori accessible space of worlds, (b) a full enough a priori grasp of the concepts in question, and (c) some nonmodal facts about the actual world. Sounds promising for modal rationalism!

More on Tim's other points later.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 08:54:33 1999
x-sender: agillies@pop.u.arizona.edu
Date: Wed, 10 Feb 1999 10:07:53 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: twins
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Dave recently wrote:

>In the subjunctive case, we're considering the standard sort of
>possibility that Kripke and others discuss. Here, I can imagine
>someone arguing that it is essential to sentient human beings (like
>me) that they be at least potentially sentient, so that the hypothesis
>that I could be a zombie is like the hypothesis (discussed by Kripke)
>that I could have descended from a different sperm and egg, etc. If
>so, the 2-intension of "me" and "DC", etc, can only pick out
>potentially conscious beings, and it's (secondarily) impossible that I
>be a zombie; counterfactual possibilities concerning a zombie should
>not be described as possibilities concerning me. (I say "potentially
>conscious" as it's quite possible that I be temporarily unconscious,
>and perhaps that I died before becoming conscious.) I'm not sure
>whether this essentialist claim is right, but it has a certain
>possibility.

This is what I had in mind re my zombie twin. The worry was whether a materialist might be able to force the zombie thought experiment to be

run along lines which would require using one's own proper name. If so, there are two options. (1) If we evaluate the scenario under PI's, then consciousness is (or arguably is) included in our PI of ourselves. (2) If we evaluate the scenario under SI's, then the name picks out the same thing in all possible worlds--and it's secondarily impossible that I could be a zombie. Either case looks problematic. Now, I don't know if any materialist would *want* (or bother) to try to come up with a way of forcing the thought experiment to make essential use of one's own proper name, but that's the sort of situation I had in mind.

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 17:09:21 1999

Date: Wed, 10 Feb 1999 17:09:18 -0800

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: twins

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Thony writes:

>This is what I had in mind re my zombie twin. The worry was whether a
>materialist might be able to force the zombie thought experiment to be
>run along lines which would require using one's own proper name. If so,
>there are two options. (1) If we evaluate the scenario under PI's, then
>consciousness is (or arguably is) included in our PI of ourselves. (2)
>If we evaluate the scenario under SI's, then the name picks out the same
>thing in all possible worlds--and it's secondarily impossible that I
>could be a zombie. Either case looks problematic. Now, I don't know if
>any materialist would *want* (or bother) to try to come up with a way of
>forcing the thought experiment to make essential use of one's own proper
>name, but that's the sort of situation I had in mind.

I'm not sure just what you have in mind by "forcing the thought-experiment to use one's proper name", but it seems to me that that couldn't really be done, since proper names are strictly speaking irrelevant to the argument. The anti-materialist needs to argue that "P -> Q" is not a priori and not necessary, where P is the complete microphysical truth about the world and Q is a phenomenal truth. They can make that case quite independently of considerations about proper names. Maybe a materialist could try to argue that "P' -> Q" is a priori or necessary, where P' is some variant on P that includes proper names, but even if that conclusion were established, it wouldn't have much bearing on the question of whether materialism is true or false.

At best, what would be going on is the same as what might go on in the case where P'' includes facts about love (say), where we take the view that love requires consciousness. Then P'' will entail some facts about consciousness, but that implies nothing at all about materialism. As long as P doesn't entail Q, then materialism is false. The entailment from P'' just tells us that materialism is false about love, too, as one would expect. Similarly, if P doesn't entail Q but P' does entail Q, where P' includes facts about DC, all that will follow is that materialism is false about DC. Which is just what one would expect if being DC essentially involves being conscious.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 10:45:57 1999
Date: Wed, 10 Feb 1999 11:45:14 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Folks,

What follows are some impressionistic "minutes" regarding today's meeting, compiled by Tim, Thony and Brad, and loosely stiched together.

Our discussion seemed to focus on two issues, broadly conceived.

(1) What are primary intensions? We spent a lot of time trying to work out whether proper names, such as London, have primary intensions, and if so, >> what are they. There was some discussion of whether 'London' and 'Londres' have different primary intensions, if they have primary intensions at all.

> >
> >> Thony argued that primary
> >> intensions aren't epistemic in this sense, they are simply functions from
> >> possible worlds to individuals or properties. There seems to be some
> >> tension between the cognitive/epistemic role primary intensions are meant
> >> to play, and their metaphysical role.

>
> It sounds odd to say that a PI is a function from worlds to
> referents, and then say that different people can have different primary
> intensions for the same terms. `Water' is supposed to pick out the
> watery stuff in any world in which it is uttered, quite irrespective (I
> take it) of what folks know about water (though, presumably, most folks
> would know which stuff was the watery stuff--but this needn't be the
> case).

>
> >
> >Perhaps related to this, there was some argument about whether
even natural kind terms (our example was "water") have primary
intensions. Brad claimed that it made sense to say "Water might not have
been H2O", but Thony argued that this only made sense to him when the
modality was construed epistemically rather than metaphysically. I (Brad)
was claiming that it was true on a metaphysical reading. If it were true,
that would motivate the notion that "water" has a primary intension (since
it is certainly *not* true that H2O might not have been H2O).

> >
>
> I guess the question that I (Thony) wanted to raise (though not at all
clearly) was
> whether the semantically salient thing that names (natural kind terms,
> etc.) contribute to sentences/discourses in which they occur is anything
> like a primary intension. I'm (Thony) not at all sure, but my intuition
seems to
> favor SI's in this respect.

>
> >
> >>
> >>
> >> This got us on to a discussion of whether it is words that are the
> >> (primary) bearers of primary intensions, or concepts. I (Tim) think
we concluded

> > that both words and concepts are meant to have primary intensions. There
> > was some discussion of whether a word has a single
> > primary intension associated with it. I (Tim) suggested that it is
not obvious
> > that this is the case: Imagine that I know that ice, liquid water and
> > steam are the same stuff. I call all three kinds of stuff "water" and its
> > part of my primary intension that all three kinds of stuff are the same
> > stuff. Imagine that my kids are only exposed to liquid water, and I don't
> > teach them that water can boil or freeze. Two questions: (A) When I talk
> > with my kids
> > about water (liquid water), are we using the same word (I guess I want to
> > say yes, tentatively); (B) do we attach the same primary intension to it
> > (I want to say no, tentatively).
> >
> > {It also seems hard to draw a hard and fast distinction between primary
> > and
> > secondary intensions, given the way kids learn words/concepts. If content
> > really has this dual structure, wouldn't one expect it to be reflected
> > more clearly?}
> >
> > But most of our discussion on this point concerned whether "London" has a
> > primary intension, and if so, what is it? We also spent some time talking
> > about the primary intentions of proper names, like 'Brad', and indexicals,
> > like "me".
> >
> > (2) Next, we talked about the debate that Thony and I (Tim) have had
regarding
> > the issue of trans-world individuation: does Thony have to be able to pick
> > out his zombie
> > twin, rather than just a zombie who is physical/functional type identical
> > to him, to get the zombie argument going. This got us on to the topic of
> > why it is important for the
> > zombie argument that I know that I am conscious. Here is how I see the
> > argument as running:
> >
> > (A) I know I am conscious
> > (B) We assume that consciousness is logically supervenient (for reductio).
> > Thus,
> > (C) Any being that is physically-functionally type identical to me will,
> > in any possible world, also be conscious.
> > (D) I can conceive of an individual that is physically-functionally type
> > identical to me, but is not conscious.
> > [Such an individual is my zombie "twin" in the sense that we are type
> > identical, but note that it need not be
> > *me* in any deep sense.]
> > (E) Conceivability entails logical possibility.
> > (F) from D and E, it follows that there is a world in which there is an
> > individual that is physically-functionally type identical to me, but lacks
> > consciousness.
> >
> > But (F) is inconsistent with (C), so (B) is false.
> >
> > I think you can run the zombie argument without making any claims about
> > the trans-world identity of individuals. **Although** you might have to
> > make commitments to the trans-world identity of properties, such as the
> > property of consciousness, and this might be problematic.
> >
>
Thony: This formulation is good, and is enough to establish that either
(B) or
> (A) is false. Now, my (Thony) thinking yesterday (and perhaps today)

was that in

> order to avoid someone concluding that (A) is false (and (B) true),
> imagining one's zombie twin becomes an exercise in considering
> counterfactual possibilities--and these require the 2-intensions.
> Further, in these subjunctive contexts, it looks like it is necessary
> that I am conscious.
>
> My main concern was not so much how/whether a materialist might be able
> to push the zombie argument into involving crucial use of *my zombie
> twin*. But, if one could make such a move, it seems that we get some
> strange consequences about conceivability and possibility.
>
>
>
> "Curious green ideas sleep furiously."
>
>

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 19:07:58 1999
Date: Wed, 10 Feb 1999 19:07:54 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Do names have primary intensions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on the "minutes".

>Our discussion seemed to focus on two issues, broadly conceived.
>(1) What are primary intensions? We spent a lot of time trying to
>work out whether proper names, such as London, have primary
>intensions, and if so, >> what are they. There was some discussion of
>whether 'London' and 'Londres' have different primary intensions, if
>they have primary intensions at all.

Actually, I discuss the 'London'/'Londres' case a bit in "The
Components of Content". Everyone should read that paper, since it
gives a lot more detail on the 2-D framework, though it is primarily
cast in terms of concepts, not words.

Do proper names have primary intensions? As I said yesterday, they
arguably don't have PIs that are universal to all users of a name. It
may be that two different users have different PIs for "London", just
as for "Bill Clinton", etc. So this might lead one to say that names
don't have PIs as a matter of "semantics", where the semantics of a
term is supposed to be universal to all users of the term.

Still, I think that it is plausible at least that any name has a PI
for any user on any occasion of use. We might put this by saying that
"name tokens" have PIs (or are associated with PIs), even though "name
types" may not. If so, we can arguably see PIs at least as part of
the semantics of an "utterance" (cf. Reimer's distinction), and

certainly as part of the semantics of an underlying thought.

Why think that name tokens have primary intensions? Basically, because like any referring word, a name (token) must have some pattern of application across epistemic possibilities. Just say I use the name "Godel". Then I can consider all sorts of epistemic possibilities about the way the actual world might be, and I can see how the name will refer with each of them.

Think of Kripke's own methodology in Lecture 2 -- we consider e.g. the epistemic possibility that the incompleteness of arithmetic was proved by a guy called "Schmidt" and stolen by a guy called "Godel" who published it, moved to Princeton, etc. Kripke notes, in effect, that if that epistemic possibility is actual, our term "Godel" refers to the Princeton guy. That's to say that the primary intension of "Godel" picks out the Princeton guy in the relevant centered world. One can do the same thing for a very wide range of worlds considered as actual (i.e. considered as epistemic possibilities): for a large numbers of such worlds, there seem to be clear facts about what our words will pick out if those worlds are actual.

We can think of the primary intension as the "reference-fixing" conditions of our terms, if we like, though actually I prefer to think of it as giving the epistemic application-conditions of our terms, or the "epistemic profile" (as opposed to the "modal profile" of the secondary intension). It's central to the way language works that any referring term has an epistemic profile (at least for a given user on a given occasion): there are facts about how it applies to different epistemic possibilities, and about how it will refer if those possibilities turn out to be actual.

How does one evaluate the primary intensions of names that one uses? Basically, take a name, e.g. "London", take a given centered world, and ask oneself "to what does the name refer if that world is actual"? Here, one considers the world as actual, i.e. considers it as an epistemic possibility: "what if the world actually turns out to be that way?". And for a very wide range of such worlds, we have clear intuitions about how the name refers. E.g., if the actual world turns out to contain XYZ in the oceans etc, then we'll say that "water" refers to XYZ. If the actual world turns out to have the Princeton guy stealing the proof from Schmidt, we'll say that "Godel" refers to the Princeton guy. If the actual world turns out to have a non-whale-eaten guy at the other end of a causal chain from our use of "Jonah", we'll say that "Jonah" refers to that guy. Etc, etc.

Summing up a primary intension in language is often difficult, and the same goes for these cases; what really matters is the function from worlds to referents, not any capsule summary. But if I were to try to get at some of the things that are involved e.g. in the PI of my name "Godel", we might try: "the guy called 'Godel' who's at the other end of a causal chain from my use of the name", or something like that. That's imperfect, as Kripke argues -- e.g. one can consider epistemic possibilities in which the guy my term refers to wasn't called "Godel" at all (I've got his name wrong, or some such). But that just means that we have to refine our view of the PI. Importantly, Kripke's own methodology here relies precisely on evaluating how the term will refer if a given epistemic possibility turns out to be actual, i.e., on evaluating the PI of the term at a world.

As for "London" and "Londres": think of Pierre's situation. For him, there are lots of epistemic possibilities. The world he thinks he is in is one with a beautiful faraway city by the name of "Londres", and

an ugly one close at hand called "London". Pierre quite reasonably says that if that epistemic possibility is actual (as he believes it is), then his terms "London" and "Londres" name two different cities: "London" picks out the faraway city and "Londres" the city close at hand. So in that centered world, the PIs of "London" and "Londres" give different results.

On the other hand, another epistemic possibility for Pierre (at least a broad epistemic possibility in the sense articulated earlier) is the actual possibility: i.e., that the people he got the term "London" from were actually referring to the same city he's living in now, and that it has both beautiful and ugly parts, etc. If Pierre were confronted with the hypothesis that this epistemic possibility is actual, he should rationally conclude that under that hypothesis, his terms "London" and "Londres" pick out the same city. So that's to say that in this centered world (Pierre's actual world!), the PIs of "London" and "Londres" give the same results.

So, Pierre's "London" PI and his "Londres" PI give the same result on some centered worlds (including the actual world), but different results on other worlds (e.g. the one he thinks is the case). So they are at least slightly different intensions overall.

How to summarize these intensions in language? Again, it's difficult and imperfect, but we might make a first attempt by saying that his "London" PI picks out "the beautiful city I've heard of under the name 'London'", and his "Londres" PI picks out "the ugly city I'm living in", or perhaps "the city I've heard of under the name 'Londres'". Getting the details right will depend on careful consideration of cases, and will depend on just how Pierre's conceptual system is set up, etc. But this gives us enough to see how the PIs might pick out the same extension in one world (the actual world) but different extensions in some other worlds.

The fact that Pierre's PIs are different precisely reflects the fact that it is epistemically possible for him that London isn't Londres. This suggests a general principle: when " $A=B$ " is a posteriori (for a user), A and B have different PIs. To see this, one can go through the following reasoning.

(1) " $A=B$ " is a posteriori;

so (2) it is (broadly) epistemically possible that A is not B,

so (3) there is some epistemic possibility in which my terms "A" and "B" pick out different things (when that possibility is considered as actual)

so (4) there is some world in which the PI of "A" and the PI of "B" yield different extensions

so (5) A and B have different PIs.

I hope all the steps here make sense. You can illustrate it by thinking of the London/Londres case, or the Hesperus/Phosphorus case.

Note that the principle here is equivalent to: when " $A=B$ " is a posteriori, " $A=B$ " has a contingent primary intension. (N.B. the primary intension of a statement is just the obvious generalization of the primary intension of a term -- in the book I call this a "primary proposition", but I now tend to use the unified terminology.) If A and B have different PIs, there's a centered world where they pick out different extensions, and in which the PI of " $A=B$ " is false. And to

say that "A=B" has a contingent PI is just to say that there's a world in which the PI is false.

This is an instance of a very important general principle:

if a statement S is a posteriori, S has a contingent primary intension.

It's fair to say that this is at the heart of the two-dimensional account of a posteriori necessity. The central claim of this account is

if a statement S is necessary a posteriori, S has a contingent primary intension and a necessary secondary intension.

That's not too hard to illustrate by working through the standard Kripke cases. I note that there are just possibly some counterexamples to this principle which someone might put forward, though they probably won't be the standard Kripke cases. We'll be discussing potential counterexamples later in the course, but feel free to suggest any now.

But anyway, one can see how all this hooks in with the need for names to have PIs. Presumably a lot of statements involving names will be a posteriori for many or all users, e.g. "London = Londres", "Cicero = Tully", etc. That's just to say that for a given user, there is the epistemic possibility that the statement is false, i.e., there are some scenarios such that if they turn out to be actual, the statement turns out false. So the very a posteriority of the statement indicates that the names in question have distinctive patterns of applications to epistemic possibilities, i.e. that they have (distinct) PIs.

That's enough for one message. Hopefully it gives useful background for thinking about issues in the rest of the minutes. More on the rest later.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 20:47:07 1999

Date: Wed, 10 Feb 1999 20:47:03 -0800

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

More on the minutes.

>> >> Thony argued that primary
>> >> intensions aren't epistemic in this sense, they are simply functions from
>> >> possible worlds to individuals or properties. There seems to be some
>> >> tension between the cognitive/epistemic role primary intensions are meant
>> >> to play, and their metaphysical role.

Hmm, I'm not sure exactly what the tension is here. Say more? On my own view, primary intensions are in a sense "epistemic", in that they involve the way our terms apply to worlds considered as epistemic possibilities, but at the same time they are perfectly respectable functions for worlds to extensions. And the worlds in question may be perfectly respectable metaphysical possibilities at the same time. E.g., the XYZ-world, with XYZ in the oceans etc, is presumably a perfectly respectable world. Considering the world as counterfactual,

our term "water" picks out H₂O in that world; considering the world as actual (as an epistemic possibility), "water" picks out XYZ.

At least, nothing in Kripke's discussion suggests that the worlds we are considering as epistemic possibilities can't all be metaphysically possible worlds in their own right. That's the centerpiece of the 2-D framework -- one space of worlds, two sorts of semantic evaluation. Somebody might want to argue that some of these epistemic possibilities don't correspond to metaphysical possibilities even in the 2-D way, but such a claim would have to rest on considerations quite distinct from Kripke's. That's something we'll be talking about later on.

>> It sounds odd to say that a PI is a function from worlds to referents, and then say that different people can have different primary intensions for the same terms. 'Water' is supposed to pick out the watery stuff in any world in which it is uttered, quite irrespective (I take it) of what folks know about water (though, presumably, most folks would know which stuff was the watery stuff--but this needn't be the case).

What exactly is odd here? I talked yesterday about some ways in which the PI of a term such as "water" might vary between individuals. Take the chemist who associates "water" only with certain chemical properties (and let's say, intends to use the term entirely without semantic deference). For him, if the XYZ-world turns out to be actual, "water" will still refer to H₂O, not XYZ. So his PI for "water" is quite different from ours. He'd be a very unusual user, of course, but it's not obvious that the story is incoherent. One might resist by suggesting that his word "water" is not really our word, and part of me would have some sympathy, but that would be a tricky row to hoe.

And without going to this sort of extreme, one can arguably find everyday cases in which two different users of a natural kind term have slightly different PIs (as witnessed by the fact that they would react to certain epistemic possibilities in different ways). E.g. a miner who is an expert on "chromium" and an airplane worker (I don't know anything about chromium, so let's pretend that it is found in mines and used in airplanes). Arguably the miner could pick out chromium as "the stuff with such-and-such superficial properties that I mine", and the airplane worker could pick it out as "the stuff with such-and-such chemical properties that I work on", or some such. Or at least, the miner might give a little more weight than the airplane worker does to superficial properties. All this is complicated by issues about semantic deference (where users intend to refer to what the community around them calls "chromium", as one gets in e.g. Burge's "arthritis" cases) but I'm taking it here that it's at least possible to use such terms without semantic deference.

I take it that Kripke's view is that natural kind terms are like names in this regard as well as in the "rigid designator" modal regard -- i.e., he thinks they don't have any canonical "mode of presentation" under which they are picked out. I.e., no universal PIs for natural kind terms. I don't think this matter is cut-and-dried, and the issues of to what extent natural kind terms are like names here is at least somewhat debatable; certainly, there is an intuition that they have some sort of descriptive "feel". It's a tricky issue.

It's true that when thinking about natural kind terms, it's easy to slip into a picture on which every user has the same PI, and I don't

think there's much harm in doing that (and it doesn't do huge violence to situation that most of us find ourselves in), as long as we recognize that it may turn out to be an idealization.

In any case, even if Kripke is right, it's not obvious that serious problems arise for us if different users have different PIs for the same terms. For most purposes that will matter for us, it might as well just be a single user's PI that we're interested in. The effect will be that what counts as (say) an explanation of "water" for us may differ from what counts for the chemist, but that won't indicate any major metaphysical difference, just a difference in the way we use the term. The general picture of metaphysical and explanation (abstracting away from surface terminology) will stay the same.

>Perhaps related to this, there was some argument about whether
>even natural kind terms (our example was "water") have primary
>intensions. Brad claimed that it made sense to say "Water might not have
>been H₂O", but Thony argued that this only made sense to him when the
>modality was construed epistemically rather than metaphysically. I (Brad)
>was claiming that it was true on a metaphysical reading. If it were true,
>that would motivate the notion that "water" has a primary intension (since
>it is certainly *not* true that H₂O might not have been H₂O).

Hmm. I guess I'd argue that natural kind terms (or term tokens) have PIs in just the same way as I argued that names have PIs: they have a distinctive pattern of application to epistemic possibilities.

Re "water might not have been H₂O": personally, I find this ambiguous between what I'd call an epistemic and a "subjunctive" reading (I prefer not to use "metaphysical" here since I think both are metaphysical in a way). Given the way the actual world is, I think Kripke's makes a pretty strong case (though not an unarguable case) that there is no counterfactual world in which water is not H₂O. If so, the subjunctive reading comes out false. But it's still the case that it's broadly epistemically possible that water isn't H₂O, i.e. there a possible world such that when that world is considered as actual, "water is H₂O" comes out false. So I think one can argue that the epistemic reading comes out true. One might argue over whether the sentence in question actually has the epistemic reading (one could argue it is intrinsically subjunctive). I personally think language is pretty flexible here, but if one has doubts, one can substitute something more clearly compatible, such as "water might not be H₂O".

I'm not sure wheter I'm agreeing with Brad or Thony here. I guess I agree with Thony that it's true on an epistemic reading, and the question of the "metaphysical" depends on just what one means by that. If one means a Kripkean subjunctive reading, then I suppose I'm inclined to agree with Kripke that the sentence is false, though I don't think the issue is entirely cut-and-dried. But as I said, I'm inclined to think that even the epistemic reading is "metaphysical", in that it involves evaluating perfectly respectable metaphysically possible worlds.

Anyway, I think that whatever one says about this, water comes out having a PI either way. All one needs for a PI is a distinctive pattern of application across epistemic possibilities (considered as actual), and that's just what we seem to find. E.g., it's epistemically possible that water isn't H₂O, i.e., there a scenario (the XYZ-world) just that when it's considered as an epistemic possibility, "water is H₂O" comes out false, i.e., the PI of "water = H₂O" is false in that scenario, i.e., "water" and "H₂O" have different PIs.

>I guess the question that I (Thony) wanted to raise (though not at all
>clearly) was
>whether the semantically salient thing that names (natural kind terms,
>etc.) contribute to sentences/discourses in which they occur is anything
>like a primary intension. I'm (Thony) not at all sure, but my intuition
>seems to favor SI's in this respect.

That's an interesting question that depends a lot on what we mean by
"semantically salient". It's true that given that PIs can vary
between users and that SIs don't, one might argue that SIs are most
central to semantics. On the other hand, when one looks at what goes
on in communication and understanding (e.g., what I learn when
someone tells me that there's water in the sink), I think it's
arguable that PIs are more central. It's a tricky issue.

>This got us on to a discussion of whether it is words that are the
>(primary) bearers of primary intensions, or concepts. I (Tim) think
>we concluded that both words and concepts are meant to have primary
>intensions.

Right. On my view, both words (or word tokens) and concepts have PIs.
The treatment in the book is largely directed at words; the treatment
in "The Components of Content" is largely directed at concepts. I
think one can argue that there is a certain sense on which the PIs of
concepts are "primary", in that the PI of a user's token of a word is
derivative on the PI of the user's corresponding concept. But that
matter isn't entirely obvious.

>There was some discussion of whether a word has a single
>primary intension associated with it. I (Tim) suggested that it is
>not obvious
>that this is the case: Imagine that I know that ice, liquid water and
>steam are the same stuff. I call all three kinds of stuff "water" and its
>part of my primary intension that all three kinds of stuff are the same
>stuff. Imagine that my kids are only exposed to liquid water, and I don't
>teach them that water can boil or freeze. Two questions: (A) When I talk
>with my kids
>about water (liquid water), are we using the same word (I guess I want to
>say yes, tentatively); (B) do we attach the same primary intension to it
>(I want to say no, tentatively).

Ah, excellent. That's a nice example of the sort of thing I was
talking about before. I think it clearly seems intuitive to say that
we're using the same word. Do we have the same PI? Actually, that's
a bit tricky. We may have different beliefs, but it certainly isn't
the case that every difference in beliefs is reflected in a difference
in PIs. What matters intuitively for a PI are the beliefs that a user
implicitly takes to be "constitutive", or a priori. It's not obvious
to me that your belief that water has three forms is one of those
beliefs. E.g., just say it turns out that the liquid around you only
has one form, and that the stuff in freezers and that comes out of
kettles is actually some entirely different chemical compounds. I
think then you'd say that "water" refers to the liquid but not to the
other things. If so, the "epistemic profile" of your "water" concept
doesn't require that water have the three forms. Arguably your PI for
"water" is more centered on the liquid stuff, just like the kids'. So
it's not completely obvious that this is a case of different PIs.
Still, it might not be too hard to turn it into such a case, or to
find other such cases in the vicinity.

>{It also seems hard to draw a hard and fast distinction between primary

>and

>secondary intensions, given the way kids learn words/concepts. If content
>really has this dual structure, wouldn't one expect it to be reflected
>more clearly?}

I'm not sure exactly what the concept-learning considerations are here. On my view, the central structure of our concepts involves PIs, as these give the conditions whereby our concepts apply to the actual world, are most associated with "core beliefs", determine the epistemic possibilities that we take to be live options, etc. SIs have a secondary role, giving the application-conditions of our concepts to counterfactual worlds, and mostly are revealed in counterfactual thinking and in subjunctive contexts. So I suppose that SIs may be revealed in kids' thinking about counterfactual scenarios ("just say I had a twin brother"), and how they describe them, etc. But probably PIs are more central, as these are what are centrally involved in the kids beliefs about the way the actual world is, about the epistemic possibilities they're prepared to entertain, etc. So I think one would probably find it easier to "read off" a PI from a kid's surface conceptual structure than an SI; and indeed, firm SIs may take longer to develop than firm PIs (given that thinking about counterfactual scenarios can be very loose). But this all brings up some very complex and arguable issues (I don't think anything I just said is obvious or uncontestable, and don't worry if it's hard to follow).

I hope this all helps to clarify some issues about the way the 2-D framework works. Please go ahead and fire away with follow-up questions, other thoughts, etc.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 22:54:19 1999

Date: Wed, 10 Feb 1999 23:53:35 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik A Herman <erikh@U.ARIZONA.EDU>

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Hi,

My main issue is Kripke's discussion of water, heat, and light (pg. 128). He says, and I agree, that "we identified water originally by its characteristic feel, appearance and perhaps taste..." But he goes on, and this is where I don't agree, to say that "if there were a substance...which had a completely different atomic structure from that of water, but resembled water in these respects, would we say that some water wasn't H₂O? I think not." He goes on to make an analogy to fools gold (fools water). My difficulty is in the amount of weight he/we are putting on physics. The reason fools gold isn't gold isn't because it is molecularly different but because the feel/appearance/(taste) part doesn't correspond to gold. My contention is that if everything were the same gold/fools gold, water/fools water than we WOULD say when we finally discovered the difference under the microscope that there are two kinds of water and two kinds of gold-- for one thing it would be impossible to determine which was the *right* one. As it turns out we understand these things in terms of the feel/appearance/taste rather than as the molecular descriptions they happen to have. Likewise, I feel the same way about heat and light. I do not think heat IS the motion of molecules or that light IS a stream of photons, rather they are the phenomenal qualities

that happen to be stimulated by molecule motion and photon streams. I realize the importance of separating phenomenal effects from physical events but we shouldn't use the same word to describe both.

Am I running into a typical problem?
Is there a name for the position I take?

More importantly, is there something big that I'm overlooking (I always feel like there is.)

Do I have it right that, in this case the primary intension would be the feel/appearance/taste? i.e. the essential qualities?

This stuff can get a little confusing.

:)

Erik H.

P.S. I don't see how consciousness could be a priori OR a posteriori -- could we make another category for it?!

Also, I can't see, especially in light of everything, how Chalmers could possibly think he's a zombie. Or any of us for that matter.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 11 06:10:05 1999
Date: Thu, 11 Feb 1999 06:09:55 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Erik H. raises an interesting point: maybe the reason we say that fools' gold isn't gold is not because it has different internal makeup, but because it has slightly different surface properties. That's an interesting and subtle issue. Things are complicated slightly by the fact that fools' gold is found in a lot of the same places we find gold. Because of that, you may well be right that if it were just the same on the surface, we'd say that there were two sorts of gold, just as we say there are two sorts of jade with different structure, jadeite and nephrite. (After all, both would be at the other end of the relevant causal chain!) But how about if there turned out to be superficially identical fool's gold on a distant planet, or if it were synthesized in a lab? If that were so, I think the chances that we'd call it "gold" would diminish, though it isn't completely obvious.

One thing we can say is that having the same intrinsic structure as the stuff we've been in contact with certainly helps get the label of "gold". Think about "carbon": certain superficially very different substances all get the label "carbon" because of their common internal structure, whereas stuff that looked like that with different structure certainly wouldn't get the name. So certainly one important factor in determining reference is having the same internal structure as the stuff we've been in contact with. So that suggests that surface qualities can't be the only factor (as Erik suggested at one point). But this doesn't imply that intrinsic structure must be the only factor, either (as Kripke suggests). The possibility remains that e.g. a disjunction of the right surface quality or the right intrinsic structure is enough to deserve the name. How to assess this?

One relevant data point involves different rare-ish metals with the same surface structure, or very similar structure (titanium and chromium, say, though I have no idea). Here, it seems that finding a metal with the same surface structure as titanium isn't enough for us to call it titanium. Titanium goes with a particular atomic structure, and that seems to be that. So one might argue that the same goes for gold. does, too.

And what would we say if a time of astronauts find a planet with the superficially identical XYZ? Water or not? Putnam and Kripke bet that we'd say it isn't water -- see "The Meaning of Meaning" for a much more extended argument for this sort of claim. I recall that I was sort of skeptical when I first heard the claim myself, but I've gradually come around. I can still see the intuition, though, and part of me can still go both ways.

>Likewise, I feel the same way about
>heat and light. I do not think heat IS the motion of molecules or that
>light IS a stream of photons, rather they are the phenomenal qualities
>that happen to be stimulated by molecule motion and photon streams. I
>realize the importance of separating phenomenal effects from physical
>events but we shouldn't use the same word to describe both.

That's another tricky one. It raises the specter of the old riddle, "if a tree falls in the forest with no-one around, does it make a sound?". To which one standard answer is to say, it makes a sound in one sense but not another. We might say it makes an "objective sound" (the air movement) but not a "phenomenal sound" (the subjective experience). And one could argue that the term "sound" is ambiguous between these. One might make the same ambiguity claim for light and heat: objective light vs phenomenal light, objective heat vs. phenomenal heat.

You want to go further and say that terms such as "heat" refer to phenomenal heat, not objective heat. One reason for questioning this might lie in the fact that we sometimes say "it's really hot in there" even though there's no-one who's actually feeling the heat (e.g. inside a furnace). Similarly we can talk about the light sent from Alpha Centauri to others star (which maybe no-one sees, etc). So it seems that objective heat and objective light are at least sometimes called "heat" and "light".

Do we sometimes use the terms for phenomenal heat and for phenomenal light. I think maybe so, though it isn't so clear. Kripke seems convinced that in a world where someone feels our "heatish" sensations but caused by something very different in the external world, we'd say there's no heat. But it doesn't seem completely unreasonable to say that there is heat in that world, that the external things are hot, etc. It's a little less clear for "light", but maybe.

If you really insist that we shouldn't use the same word for both, my own feeling is that there is slightly more of a case for using the term for the objective kind than for the phenomenal kind. But of course it is ultimately a terminological issue. One of my fundamental principles in philosophy is that nothing of vast importance ever really turns on a terminological issue. But of course the way the terminology of natural language works is of some interest in its own right.

>Am I running into a typical problem?
>Is there a name for the position I take?

>
>More importantly, is there something big that I'm overlooking (I always
>feel like there is.)
>
>Do I have it right that, in this case the primary intension would be the
>feel/appearance/taste? i.e. the essential qualities?

I can't think of a name, offhand, but I think your view isn't entirely unusual. Some people have argued that the reference of our terms such as "heat" has gradually shifted from the phenomenal kind to the objective kind as science has progressed: i.e., we have gradually "objectified" the term. So when Goethe argued that light couldn't be explained by science, it wasn't so much that he was proved wrong as that the subject was changed.

Either way, the PI of "heat" etc will involve the appearance. If we think "heat" refers to phenomenal heat, the PI will just pick out that particular sort of sensation in all worlds, as will the SI. If (like Kripke) we think it refers to objective heat, the PI in a given world will pick out the external property that causes those sensations around the center of the world. The SI will pick out the *actual* external property (the motion of molecules, given the way the 'actual world turned out) in all worlds. (So the PI can be roughly summarized as "the cause of heat sensations", and the SI as "the motion of molecules".)

>P.S. I don't see how consciousness could be a priori OR a posteriori --
>could we make another category for it?!

I take it you mean knowledge of consciousness. I'm sympathetic with your point here -- there's certainly some sense in which we know about our consciousness non-empirically and certainly, so it can seem a priori. And some philosophers have argued that it is a priori. But I think for a number of reasons it is best to see it as a posteriori. One's knowledge of consciousness is knowledge in virtue of experience, which is just what a posteriori knowledge involves. It isn't a "truth of reason" in any clear sense. One can argue that phenomenology and introspection are just as "empirical" as external observation: one measures the first-person data, the other the third person data. And finally, a lot of philosophical issues make more sense if we see this as a posteriori. So I think there's a good case for putting it in this category, even if it has a somewhat different feel from the paradigm cases of a posteriori knowledge.

> Also, I can't see, especially in light of everything, how Chalmers
>could possibly think he's a zombie. Or any of us for that matter.

I don't think I'm a zombie! All I said is that it's not a priori that I'm not a zombie, and that it's 1-possible that I'm a zombie. That's entirely compatible with the claim that I am certain that I'm not a zombie. It just turns out that I am certain of something that is a posteriori. That seems OK to me, but if there's a problem, I'm interested to hear it. As for you, I'm not certain you're not a zombie, but I think you're not, and arguably I know you're not. But still, it is certainly a broad epistemic possibility that you're a zombie; i.e., it isn't ruled out a priori.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 12 14:17:22 1999
Date: Fri, 12 Feb 1999 15:16:38 -0700
Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Timothy J Bayne <bayne@U.ARIZONA.EDU>

Subject: Proper Names

Comments: To: David Chalmers <chalmers@LING.UCSC.EDU>

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

A thought on proper names. . .

There is an interesting tension between two aspects of the proper names that we use. On the one hand, we want a proper name to be unique to the individual whose names it is. (This fact, or rather the absence of this fact, is the basis for Monty Python's Australian Philosophers sketch, in which everyone is called Bruce.) On the other hand, individuals do share proper names, lots of people are called Bruce. It is interesting to note that there is anecdotal evidence that kids resist calling another individual by a proper name that they have already attached to an individual: If my sister is called Jane, then it's wrong to call this other person Jane as well. They seem to over-extend the principle that proper names aren't like nouns and adjectives, they are meant to apply to unique individuals.

How does this relate to the "meaning" of proper names? Consider a conversation in which I am talking about my sister Jane, and not my friend Jane. Or, better, a conversation in which I'm talking about my sister Jane, and someone interrupts to tell me that "Jane wants to talk to me", and I interpret the second "Jane" to be about my friend. In order to keep the references of the various tokens of Jane straight, I presumably have to bind some kind of descriptive content to them. At some level, don't I have to be thinking "Jane-my-sister" or "Jane-my-friend", or something with *some* descriptive content?

The precise nature of the descriptive content needed here is not obvious, but it is clear that a traditional descriptivist suggestions won't work: "The person that everyone around here calls *Jane* won't work", because there are lots of people around here called Jane.

In other words: in general, the meaning of a proper name can't just be its referent, 'cos most proper names have a number of referents. One could respond to that by saying that *Jane* in one context means something different from *Jane* in another context, because it is used to refer to a different person. Thus, Jane/Jane would be just like bank(financial)/bank(river). But the analogy with "bank" seems to me to play into the descriptivist's hands. When I say "I saw you at the bank yesterday" I know that I either mean river bank or financial bank 'cos of the descriptive content I associate with the token. Unless I also associate descriptive content with 'Jane', how do I (or my interlocuters) know which Jane I am talking about?

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From bayne@U.Arizona.EDU Fri Feb 12 14:17:19 1999
Date: Fri, 12 Feb 1999 15:16:38 -0700 (MST)
From: Timothy J Bayne <bayne@U.Arizona.EDU>
To: David Chalmers <chalmers@LING.UCSC.EDU>
cc: MODALITY@LISTSERV.ARIZONA.EDU
Subject: Proper Names
Status: RO

A thought on proper names. . .

There is an interesting tension between two aspects of the proper names that we use. On the one hand, we want a proper name to be unique to the individual whose names it is. (This fact, or rather the absence of this fact, is the basis for Monty Python's Australian Philosophers sketch, in which everyone is called Bruce.) On the other hand, individuals do share proper names, lots of people are called Bruce. It is interesting to note that there is anecdotal evidence that kids resist calling another individual by a proper name that they have already attached to an individual: If my sister is called Jane, then it's wrong to call this other person Jane as well. They seem to over-extend the principle that proper names aren't like nouns and adjectives, they are meant to apply to unique individuals.

How does this relate to the "meaning" of proper names? Consider a conversation in which I am talking about my sister Jane, and not my friend Jane. Or, better, a conversation in which I'm talking about my sister Jane, and someone interrupts to tell me that "Jane wants to talk to me", and I interpret the second "Jane" to be about my friend. In order to keep the references of the various tokens of Jane straight, I presumably have to bind some kind of descriptive content to them. At some level, don't I have to be thinking "Jane-my-sister" or "Jane-my-friend", or something with **some** descriptive content?

The precise nature of the descriptive content needed here is not obvious, but it is clear that a traditional descriptivist suggestions won't work: "The person that everyone around here calls **Jane** won't work", because there are lots of people around here called Jane.

In other words: in general, the meaning of a proper name can't just be its referent, 'cos most proper names have a number of referents. One could respond to that by saying that **Jane** in one context means something different from **Jane** in another context, because it is used to refer to a different person. Thus, Jane/Jane would be just like bank(financial)/bank(river). But the analogy with "bank" seems to me to play into the descriptivist's hands. When I say "I saw you at the bank yesterday" I know that I either mean river bank or financial bank 'cos of the descriptive content I associate with the token. Unless I also associate descriptive content with 'Jane', how do I (or my interlocuters) know which Jane I am talking about?

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy

University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 12 19:52:56 1999
Date: Fri, 12 Feb 1999 20:51:59 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: RE Proper Names
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

With respect to my previous comment about proper names. . . .(Re)reading the preface to **Naming and Necessity** tonight I was reminded that Kripke discusses just this point, and seems to think that it's of no real relevance. I'm not persuaded that he's justified in dismissing it in the way he does. He seems to think that you just bring in pragmatics or something to work out which Aristotle you're talking about, i.e. to work out which proposition you mean to express, and then plug his rigid designation account into the proposition that one has isolated. Well, maybe, but maybe not. But it's not obvious to me that we can simply dismiss the issue of homonymous names to speaker meaning or pragmatics or something. I don't have an argument here, but Kripke's discussion seems to me to be a little underwhelming.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 12 21:51:49 1999
Date: Fri, 12 Feb 1999 21:51:41 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Proper Names
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Lots of interesting issues about multiple uses of the "same" name in Tim's message. Starting from the end of the message:

>In other words: in general, the meaning of a proper name can't just be its
>referent, 'cos most proper names have a number of referents. One could
>respond to that by saying that **Jane** in one context means something
>different from **Jane** in another context, because it is used to refer to a
>different person. Thus, Jane/Jane would be just like bank(financial)/bank
>(river). But the analogy with "bank" seems to me to play into the
>descriptivist's hands. When I say "I saw you at the bank yesterday" I know
>that I either mean river bank or financial bank 'cos of the descriptive
>content I associate with the token. Unless I also associate descriptive
>content with 'Jane', how do I (or my interlocutors) know which Jane I am
>talking about?

Hmm. The view that the meaning of a proper name is just its referent is very popular in contemporary philosophy (held e.g. by direct reference theorists), perhaps even so popular as to be conventional wisdom. I think those who hold this view will respond by saying that the two instances of "Jane" here are simply two different names. One name means X, and one name means Y. It's just that somewhat

coincidentally, the two names are spelled and sound the same, or something like that. See for example Kripke's discussion on pp. 7-8, where he suggests that "uses of phonetically the same sounds to name distinct objects count as distinct names".

Of course one might object by saying there's an intuitive sense in we say that Jane 1 and Jane 2 "have the same name". Here a direct reference theorist (and others) might respond e.g. by saying that there are two ways of individuating names, either by their phonetics (etc.) or by their meaning; and that while phonetic individuation sometimes plays a role in ordinary practice, it is meaning-individuated name types that correspond best to the deep structure of language, or that matter most in philosophy, or something like that. As usual the issue isn't cut-and-dried, but I think the direct reference theorist has some resources here.

Your objection in turn raises a really central issue in the debate over the semantics of names: how do I know the referent of my name? The direct reference theorist might say, well, you know that the referent of "X" is X, but it might be objected that this is trivial and unhelpful. In practice something more seems to be involved in the way in which we identify the referent of our names, i.e. the way in which we come to know who or what the referent is. This seems closely tied to what we might call the "mode of presentation" of the name. The direct reference theorist may well accept that any use of a name has a mode of presentation associated with it by the speaker, or at least some way of identifying the referent, but they will deny that the mode of presentation is part of the "meaning" of the name. It might merely be something in the psychology of the speaker, for example, rather than a truly semantic phenomenon. (Often, a direct reference theorists will sharply separate "semantic" and "cognitive" issues.) Others (especially description theorists) prefer to think of the mode of presentation of a name as something in the realm of meaning and semantics.

One consideration in favor of the direct reference theorist is the point (which we've already noted a couple of times) that different users of the same names may well have different associated modes of presentation. E.g., Hillary's and my use of "Bill Clinton". One might then say it follows automatically that the mode of presentation can't be part of semantics, if the semantic features of a term are common to all users of a term, as it is often thought they must be. In response, it might be argued that the mode of presentation is at least part of "utterance meaning" rather than "sentence meaning", or something like that, where the former can be specific to a single use or user of a term, where the latter is required to be universal. Here again Marga Reimer's work on distinctions like these is relevant; I don't have the relevant papers with me (and may be misremembering them a bit), but fortunately Marga has just joined the mailing list -- any contribution is welcome!

At some point these debates about what truly counts as "meaning" and "semantics" become somewhat terminological (that's another of Marga's themes, I think), but the substantive issue remains of whether there are modes of presentation associated with names (whether or not they are part of semantics), and if so what they are. Tim in effect argues that there must be something like this, in order that we can know what we're talking about, and I'm sympathetic. One possibility is that these modes are descriptions; that would make things easy, as we could know that the referent is whoever satisfies the descriptions. But maybe there are other possibilities, e.g. that the modes of

presentation are causal chains linked to our tokens, and to identify a referent we follow down the causal chain associated with the tokens (given that we somehow know which causal chain to start following).

My own view, you won't be surprised to hear, is that these modes of presentation are precisely primary intensions. It's the case that any name (as used by a user) has its own epistemic application conditions, i.e. conditions under which it applies to epistemic possibilities, under the hypothesis that those possibilities turn out to be actual. The fact that users know what they're talking about corresponds to the fact that users have the conditional ability to identify the referent of a name given sufficient empirical information about the actual world -- that is, they have the ability to evaluate a name's primary intension at a world. So they have "cognitive access" to their names' primary intensions, and in a sense the primary intension represents the "cognitive content" of the name as used by them. And at the same times, the primary intension determines reference.

The primary intension needn't be a description in any obvious sense, but as we've seen its plausible that it will always have some associated descriptive content (core beliefs, etc).

How does all this relate to the dual-name case? I think the issue Tim raises (about knowing what we're talking about) applies even to cases without ambiguity, but it comes up in a sharp form here. When I use the names "Jane 1" and "Jane 2" (so to speak), I know (or believe) that these refer to different people. This reflects itself in my ability to identify the referent of each, given relevant information about the external world. So in many/most/all epistemic possibilities considered as actual, I can determine the separate referent of the names. This is just to say that I have cognitive access to quite different primary intensions for each name.

The question then arises about just what these primary intensions involve at the psychological/cognitive level; i.e., what is it that grounds my cognitive ability to identify the referents? The description theorist will say that it is the descriptions I associate with the name: in effect, the primary intension is given by some core or cluster descriptions. Someone else might say it is given by a reactive disposition of some kind, not necessarily mediated by descriptions. My own view is that a primary intensions needn't be represented as a description per se, but that nevertheless it will correspond to what we might think of as "tacit descriptive knowledge" about the referent. E.g., even if we identify the referent (with Kripke) by following a causal chain from a token, our very ability to do this suggests a tacit knowledge that the referent is at the other end of the causal chain, which one might think of as tacit descriptive knowledge about the referent in some sense. So on this view, there will at least be some sort of loose link between primary intensions and tacit descriptive content and the cognitive level.

Applying all this back to the Jane cases, the distinct cognitively accessible primary intensions for the two names suggests that there must at least be some sort of different descriptions tacitly associated with each name, that will emerge in the way reference is identified. Just what sort of descriptions they are will depend on the details of the case. I think one is likely to say something quite different depending on whether one is acquainted with the referent in question or has merely heard about them and picked up the name. In the case of acquaintance (e.g. the sister/friend case), I take it one will have tacit descriptive "immediate" knowledge of the people in

question and that some of this will be crucial to reference fixation (e.g., knowledge of their roles vis-a-vis you).

In the case of non-acquaintance (say I have names for David Hilbert the mathematician and David Hilbert the philosopher, though I haven't met either), we'll have deference to a community of prior users of the name, and things get a bit subtler. One might try using surface descriptive knowledge (e.g. of career) to disambiguate, but I'm not sure this gets to the heart of things, as arguably one could be wrong about those things, and there are cases where one has little such knowledge. I think it's better to note that one knows (at least tacitly) that there are two names here (or one knows that there are at least potentially two people), and one knows that one picked up each name from the community in a slightly different manner. So in each case, one fixes reference by deferring to the community, but in different ways. Maybe for one I defer to the mathematical community (where I heard the name) and the other to the philosophical community. Or maybe I don't remember much at all about which community or users I got the names from, but still, I implicitly defer in each case to "whatever users I picked up *this* name from", as opposed to "whatever users I picked up *this* name from" for the other (in effect demonstrating the different name tokens). That arguably does enough to get started on the different causal chains by which reference through deference is fixed.

Anyway, I think you're right that the ability to refer differently with the two names (or name tokens) suggests that we have some sort of tacit descriptive knowledge associated with each that reflects our ability to do so (though in the last case above, the difference in the knowledge is pretty thin). Just whether this difference in associated descriptive content (and difference in primary intension, etc) is enough to refute the direct reference theorist isn't quite clear, as they still have the option of saying that it's a matter of psychology (say), not semantics, and that the true semantics of the name still just involves the referent. But in any case it suggests that there is something important going on in the vicinity.

This is already too long, and I think we've covered most of the bases now, but a few quick notes on the other bits.

>How does this relate to the "meaning" of proper names? Consider a
>conversation in which I am talking about my sister Jane, and not my friend
>Jane. Or, better, a conversation in which I'm talking about my sister
>Jane, and someone interrupts to tell me that "Jane wants to
>talk to me", and I interpret the second "Jane" to be about my friend. In
>order to keep the references of the various tokens of Jane straight, I
>presumably have to bind some kind of descriptive content to them. At some
>level, don't I have to be thinking "Jane-my-sister" or "Jane-my-friend",
>or something with *some* descriptive content?
>

>The precise nature of the descriptive content needed here is not obvious,
>but it is clear that a traditional descriptivist suggestions won't work:
>"The person that everyone around here calls *Jane* won't work", because
>there are lots of people around here called Jane.

I hope I've mostly covered this above. One might argue that the mere existence of the two names just suggests that you have two different mental tokens (say) "Jane 1" and "Jane 2", and some sort of tacit rule for figuring out when someone's "Jane" talk corresponds to one or other symbol (e.g., your family speaking -> Jane 1, your friends speaking -> Jane 2). But the fact that you have the ability to know

who you're talking about in a deeper sense, i.e. the ability to identify the referent of each in the world, suggests as above (at least to me) that you must at least have some sort of deeper tacit descriptive content associated with each.

You're right that the nature of the descriptive content isn't obvious. It's a nice example to refute one description (the one you give) that description theorists sometimes offer. It seems that one would need to appeal to either (a) different bits of "first-order" descriptive knowledge (e.g. in the acquaintance case), or (b) different bits of "metalinguistic" descriptive knowledge (in the deference case). And as you say (in effect), the metalinguistic knowledge can't be specified just by using the word "Jane". One will either have to use knowledge of the different communities or users that one picked up the name from, or at least, the knowledge that each token is part of a separate chain of uses, and that each token is causally connected to a point where you picked it up from somewhere. The last case is an example of the very thin "tacit indexical causal descriptive knowledge" that we talked about above, but I guess it is enough to enable one to pick out the different referents, in the case where one has very thin knowledge associated with the name.

The general question of what descriptions a description theorist should appeal to to handle the Kripke cases (e.g. the Feynman, Godel, Jonah cases) and others is a really interesting one that I hope we get a chance to spend some time on at some point. Arguably there are resources that Kripke doesn't consider. His discussion on pp. 160-162 is worth looking at. He concedes (as in footnote 38) the possibility of a relatively trivial "causal descriptivism" that encapsulates the whole causal theory of reference into a description -- not exactly a strong form of descriptivism, but maybe enough for some descriptivist purposes. There are also metalinguistic options to think about, of which he considers some but arguably not the best.

Interestingly he alludes (on p.162) to one of the better of these, "Let 'Glumph' denote the man called 'Glumph' by the people from whom I got it (whoever they are)", but only in conjunction with the "trivial" causal description above. One could argue that this is better than any of the metalinguistic options he considers explicitly, although it is still imperfect -- e.g. it might turn out that you misheard the name, so that you're the only one who uses 'Glumph' but you still refer. But then one might try "the person I heard of under the name 'Glumph'", or "the person referred to by the people from whom I got the name 'Glumph', under the name causally responsible for my use of 'Glumph'". And maybe bring in "this very name 'Glumph'" to get around the ambiguity problems Tim raises above.

It gets tricky, of course, but then the description theorist never said there has to be a universal description formula for all names. Arguably, going for a "cluster" description with one of the metalinguistic descriptions just mentioned as one of the components of the cluster (probably the central component) can handle almost all of the Kripke-style cases and other cases of deference, and can go along way to grounding something like a causal theory of reference in something like a description theory. But that's a long story. Personally one of the reasons I like the primary-intension framework is that it bypasses this need to come up with specific descriptions, which can obviously get very complex quite quickly. Nevertheless I think there are ways in which one can use the framework to lend support to an attenuated version of the description theory.

--Dave.

P.S. I just got this.

>With respect to my previous comment about proper names. . . .(Re)reading
>the preface to **Naming and Necessity** tonight I was reminded that Kripke
>discusses just this point, and seems to think that it's of no real
>relevance. I'm not persuaded that he's justified in dismissing it in the
>way he does. He seems to think that you just bring in pragmatics or
>something to work out which Aristotle you're talking about, i.e. to work
>out which proposition you mean to express, and then plug his rigid
>designation account into the proposition that one has isolated. Well,
>maybe, but maybe not. But it's not obvious to me that we can simply
>dismiss the issue of homonymous names to speaker meaning or pragmatics or
>something. I don't have an argument here, but Kripke's discussion seems to
>me to be a little underwhelming.

I think you're right that Kripke doesn't get at all the central
issues. He mostly says that this issue is irrelevant to issues
concerning rigidity, and I think he's more or less right about that.
But arguably it is important for issues about the fixation of
reference, e.g. for issues about causal theories and descriptive
theories of reference-fixing, of the sort he describes in Chapter 2.
There has to be some story about how the different tokens get their
distinct referents, and as we've seen, this isn't entirely trivial.

Incidentally I heard someone a while back give a paper with the
thesis, contra received wisdom but in sympathy with your view, that
all uses of name tokens such as "Jane" are in fact tokens of the same
name. So there is just one name "Jane", one name "Paul", one name
"David Hilbert", etc. She addressed the problem of how the name gets
to refer differently in different cases by arguing that names are
implicit demonstratives -- in effect one is saying "that Jane", "that
David Hilbert", etc, with different intentions to demonstrate in each
case. A demonstrative like "that" can have different referents in
different contexts and with different intentions, but it's still one
word. Same on this view for "Jane". I'm not sure the view is
completely unproblematic, but it's interesting.

From owner-modality@LISTSERV.ARIZONA.EDU Sat Feb 13 10:41:56 1999

Date: Sat, 13 Feb 1999 11:40:43 -0700

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Rachael J Parkinson <rachaelp@U.ARIZONA.EDU>

Subject: Re: Proper Names

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

I have some questions about the Kripke reading- specifically on pages
123-133.

On page 123 Kripke acknowledges that we could be mistaken about gold
having the atomic number 79. But he argues that given that gold does have
the atomic number 79, something could not be gold without having the
atomic number 79. Kripke makes light of Kant's mistaken assertion that
gold is a yellow metal- but it seems that it is **possible** that just as we
found out that all gold isn't yellow, we may find out that we are mistaken
about gold having the atomic number 79. Kripke argues that we can pick
gold out in all possible worlds by its atomic number. But if he was
arguing 300 years ago, perhaps he might have made a similar argument- that
gold is a yellow metal in all possible worlds.

This brings up the question of what properties are essential. I
think the answer is more difficult to pinpoint in cases like 'what makes a

tiger a tiger' then for 'what makes gold, gold?' On page 121, Kripke seems to suggest that the external appearance of a tiger is not essential - we must look at its internal structure to determine whether it is a true tiger. My question is two-fold. First, if we were to discover a microscopic virus that happened to have the same internal structure as a tiger- would we recognize it as a tiger? Second, Kripke asserts that certain properties like quadrepedal, tawny yellow, and carnivorous are not essential to tigers, we might find that all these properties are optical illusions, but couldn't we discover the same of the internal structure of tigers?

I guess I have a problem with how to identify natural kinds, particularly in reconizing which properties are essential and which are not. In light of past mistakes (like gold is a yellow metal) I am not sure that we can ever know that we have actually discovered the essence of a natural kind, or what picks it out in all possible worlds. Perhaps you could say, given *our* concept of gold, what picks it out for us is the atomic number 79. But if it seems logically possible to be wrong about gold always having the atomic number 79 in this world, then why couldn't we be wrong about it in other worlds?

As our science develops we discover that our original theories about light and heat were mistaken. I'm not sure how this works- but it seems like if enough of the data (what we recognize as heat and light) don't match up with the theory- then we find a new one. It is still possible to discover that heat is not just the motion of molecules- I suppose that to accomplish such a discovery we would have to be using our primitive notion of heat by which we identify it as a characteristic effect produced on our nerve endings. If heat just *is* motion of molecules then it looks like instead of reformulating the theory- we would have to call that other kind of thing that produces a hot sensation in us- something other than heat. It just doesn't seem that "heat is motions of molecules" or "water is H₂O" are as necessary as "a bachelor is an unmarried man." What do you think?

-Rachael

From owner-modality@LISTSERV.ARIZONA.EDU Sat Feb 13 17:03:29 1999

Date: Sat, 13 Feb 1999 17:52:44 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Laurence A James <laj@U.ARIZONA.EDU>

Subject: Re: Proper Names

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Hi,

I would like to respond to some of Rachael's questions. First, you mention that Kripke doesn't deem properties such as quadrepedal and tawny yellow essential to tigers. I think the reason for this, is not so much that he is worried about optical illusions, but rather more about cases where a tiger is missing a leg (or about one of those white tigers for instance). Both of these types of tigers are still tigers regardless of their minor physical defects. Secondly you mentioned the possibilty of finding a virus that might have the same internal structure as a tiger, and asked whether or not this should be a tiger too? However, I think Kripke was being a little more direct when he mentioned internal structure. I think he was referring to something like the tiger's DNA structure. And, as far as i understand genetics, it will be quite unlikely that a virus would have the same genetic structure (i.e., DNA) as a tiger.

As for Gold having atomic number 79. This is an essential property of Gold. Gold and only gold has this atomic number. Now, we might find something that looks similar to gold, infact almost identical, but I don't see the reason why we would have to call it gold. I think that Kripke

makes a strong claim here - that for something to be Gold, no matter whether in this world or a possible one, it must have the atomic number 79. Similar looking things with different atomic numbers will not be Gold. Perhaps as an illustrative example, consider Platinum and White Gold. I assume that there is a common distinction between these two. We all agree that they are different, but Platinum and white Gold share many of the same properties. They are both pliable and malleable, they both have a silvery color, etc..

One last thing you mentioned was that it did not seem like "a bachelor is an unmarried man" and "water is H₂O" have equal degrees of necessity. You had remarked that it seems in fact that water is H₂O does not seem to be as necessary as being a bachelor and unmarried. Perhaps, this comes from the fact that bachelor, simply entails unmarried man, but it doesn't seem to be the same for water (and H₂O). However consider the following: Water, white Vinegar, and Parafin. All three look the same. They are all liquid, they all evaporate, they can all make you wet, etc. Now it is clear that these three are different. I think we can see Kripke's point here. When water was found and initially dubbed Water, the liquid took that name necessarily (given a correct dubbing ceremony by competent speakers). Given that the liquid H₂O took the name water necessarily, and that that liquid was H₂O, water is now necessarily H₂O, just like Gold has atomic number 79. So, as I think Kripke would say, both the bachelor and example have the same degrees of necessity. One further thing to think about, and perhaps an issue that makes this a little unclear, is that the statement "water is H₂O" is necessary a posteriori, while that of a "bachelor being an unmarried man" is necessary a priori. This might be why it does not seem like the two statements have the same degree of necessity, but, I think that Kripke argues for these two modes of 'coming' to the necessary (a posteriori and a priori) quite forcefully.

Anyways, I hope these comments have helped a little.

-Larry

From owner-modality@LISTSERV.ARIZONA.EDU Sat Feb 13 21:09:40 1999
Date: Sat, 13 Feb 1999 21:09:35 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Epistemic and modal issues
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Rachael raises some interesting issues re gold. To address them properly, we have to (as always) sharply distinguish Kripke's modal and epistemic points. His epistemic points concern what we "could find out" about gold in the actual world. His modal point concerns the nature of gold across possible worlds, given that it is a particular way in the actual world. The "gold" discussion in Lecture 3 is somewhat tricky because he uses different aspects of the case to make both modal and epistemic points, but I think one can still separate them.

Maybe in this message, I'll take the opportunity here to say some general things about the distinction between the modal and epistemic points and about the structure of Kripke's discussion, in case it's useful for some people. This will also tie in to the 2-D framework. Then in the next message I'll try to address Rachael's discussion.

It's sometimes tempting to run Kripke's modal and epistemic points together, because both make arguments against description theories of meaning. But the arguments concern very different issues. The modal

point generally argues that for a name or natural kind term N and a description D, it is not *necessary* that N is D (there is a possible world in which N is not D). The epistemic point usually argues that it is not *a priori* that N is D (we could find out that N is not D). As we've seen, the points are quite separable, e.g. by considering cases such as "Jack the Ripper" or "Hesperus" where arguably the modal point holds but the epistemic point doesn't.

In Kripke, the modal point is tied to the thesis that names and natural kind terms are rigid designators, picking out the same thing or stuff in all possible worlds. The epistemic point is tied to the causal theory of reference, holding that our names pick out their referent in the actual world not by description but by a causal chain. These two main themes in Kripke -- rigid designation and the causal theory of reference -- are easy to run together but should be kept well apart.

How is all this reflected in the structure of Kripke's book? He goes back and forth a bit, but in essence, the central parts of Lecture 1 are devoted to modal issues and rigid designation, the central parts of Lecture 2 are devoted to epistemic issues and the causal theory, and Lecture 3 has some of each. In more detail (though still imperfectly), I'd divide things up as follows:

Lecture 1

pp. 22-39: preliminaries (names and descriptions, apriority, necessity)
pp. 40-63: modal issues (rigidity: Nixon, meter stick, Aristotle, Hesperus)
pp. 64-70: general issues re description theory (mostly epistemic)

Lecture 2:

pp. 71-73: general issues re description theory (mostly epistemic)
pp. 74-78: modal issues (recap: Aristotle, Hitler)
pp. 79-97: epistemic issues (non-descriptive reference: Feynman, Godel, etc)
pp. 97-105: modal issues (necessity of identity, Hesperus=Phosphorus, etc)

Lecture 3:

pp. 106-116: modal issues (necessity of identity and of origins, essences)
pp. 116-123: epistemic issues (natural kinds: gold, tigers)
pp. 122-133: modal issues (natural kinds: gold, water, heat, etc)
pp. 134-143: general considerations (both epistemic and modal)
pp. 144-155: mind/body problem

OK, now what about his discussion of gold, in particular. As suggested above, the first part of the discussion (esp. pp. 116-119) focuses on epistemic issues, and the second part (esp. pp. 123-125) focuses on modal issues.

In the first part, he argues epistemically (contra Kant) that it is not a priori that gold is a yellow metal. Here he argues that we could find out (much to our surprise) that gold isn't a yellow metal, e.g. because there was a massive optical illusion. In the second part, he argues that given that gold has atomic number 79 in the actual world, it is necessary that gold has atomic number 79. Here he argues that in a counterfactual situation in which the ubiquitous yellow metal didn't have atomic number 79 (e.g. because it was all a yellow pyrite), it wouldn't be gold.

Note that both these methods of argument consider hypothetical scenarios, but that we think about the scenarios in very different

ways. In the first (epistemic) argument, we take a scenario (the optical illusion scenario) and consider what we would say if we found out that it was actual. In the second (modal) argument, we take a scenario (the ubiquitous pyrite scenario) that is explicitly counterfactual, and we consider how to describe that counterfactual scenario. Here there is no thought of "what to say if this is actual"; rather, we say "given that the actual world is as it is (with gold = element 79), how do we describe the counterfactual scenario".

This difference can be naturally translated into the 2-D framework, so I'll spend a few paragraphs doing that, to help get a feel for that framework.

To put things in the 2-D terminology, the first argument just mentioned requires us to *consider a scenario as actual*, and the second argument requires us to *consider a scenario as counterfactual*. Considering a scenario as actual involves considering it as a broadly epistemic possibility (a way the actual world may turn out to be); considering a scenario as counterfactual involves considering it as a "subjunctive" possibility (a way a world could be, given that the actual world is fixed).

Kripke's epistemic arguments generally involve considering a scenario (e.g. the optical illusion scenario) as actual and thinking about how to describe it. Kripke's modal arguments generally involve considering a scenario (e.g. the ubiquitous pyrite situation) as counterfactual and thinking about how to describe it. It's the former that's relevant to what is a priori, and the second that is relevant to what is necessary, in Kripke's sense.

Taking epistemic arguments first:

Kripke often wants to argue that some statement ("N is D", "Godel proved the incompleteness of arithmetic", "gold is a yellow metal") is not a priori. To do this, he finds a scenario that he considers as actual (considers as an epistemic possibility), and argues that the statement is false when that scenario is considered as actual (he puts this by saying that if the scenario turned out to be actual, we'd say that the statement is false). So he argues that "gold is a yellow metal" is not a priori by arguing that it comes out false when the optical illusion scenario is considered as actual.

This illustrates the general principle: if a statement is a priori, it is true in all scenarios considered as actual. Or to go all the way into 2-D terminology: if a statement is a priori, it has a necessary primary intension.

The primary intension of a statement is defined as the function that takes a possible world and returns the truth-value of the statement in that world when the world is considered as actual. When we think about an epistemic possibility and think about what to say if it is actual, we are just evaluating the primary intension of a statement there. So when Kripke is arguing that a statement would come out false if a given scenario (e.g. the proof stealing scenario, or the optical illusion scenario) is actual, he is in effect establishing that the primary intension of the statement is false in that scenario. So the statement has a contingent primary intension, which corresponds to the fact that it is a posteriori.

Doing this over again with modal arguments:

Kripke often wants to argue that some statement ("N is D", "Hesperus

is the evening star", "stick S is a meter long") is not necessary. To do this, he finds a scenario that he considers as counterfactual, and argues that the statement is false when that scenario is considered that way. So he argues that "Hesperus is the evening star" is not necessary by arguing that it comes out false when a given scenario (e.g. Venus was hit by a comet thousands of years ago) is considered as counterfactual.

This illustrates the general principle: if a statement is necessary, it is true in all scenarios considered as counterfactual. Or to go all the way into 2-D terminology: if a statement is necessary, it has a necessary secondary intension.

The secondary intension of a statement is defined as the function that takes a possible world and returns the truth-value of the statement in that world when the world is considered as counterfactual. When we think about a possibility "subjunctively" and think about how to describe it (given that the actual world is fixed), we are precisely evaluating the secondary intension of a statement there. So when Kripke is arguing that a statement is false of a given counterfactual scenario (e.g. the Venus-comet scenario, or the pyrite scenario), he is in effect establishing that the secondary intension of the statement is false in that scenario. So the statement has a contingent secondary intension, which corresponds to the fact that it is necessary.

>From all this, we can see the crucial point that Kripke's epistemic points always (in effect) concern the primary intensions of our statements and terms, and that his modal points always (in effect) concern the secondary intensions of our statements and terms. For example, the modal thesis that names are rigid designators is in effect the thesis that names' secondary intensions don't correspond to descriptions, but pick out the same individual in all worlds. The epistemic argument against the descriptive theory of reference and for the causal theory in effect corresponds to the thesis that names' primary intensions aren't given by descriptions, but by picking out whatever lies at the other end of certain causal chains.

OK, now back to the gold discussion. In effect, his epistemic point re gold (that it isn't a priori that gold is a yellow metal, e.g. because of the optical illusion scenario) is intended to count against any descriptive theory of reference-fixing for gold, and for some sort of causal account instead. His modal point re gold (that it necessarily has atomic number 79, e.g. by considering the pyrite scenario) is in effect intended to establish that a term like gold is in some sense a rigid designator (picking out the same stuff across counterfactual worlds).

The theses are separable, as usual. In effect, the first concerns the primary intension of "gold" (it picks out what's at the end of a causal chain, not what satisfies a description), and the second concerns its secondary intension (it picks out a constant atomic kind, not what satisfies a description). In this way, he aims to show that natural kind terms are in *both* respects like names. Causal reference-fixing, and rigid designation.

Corresponding to these distinct theses, we get somewhat different results for the primary and secondary intensions of "gold", as suggested above. To illustrate the difference explicitly, consider a possible world where there are superficially goldy pyrites everywhere in the mines and such, and nothing with atomic number 79 there. As Kripke notes in effect, when we consider that as a counterfactual

scenario (with the actual world fixed), "there's gold in the mines" comes out false there, and "gold has atomic number 79" comes out true. (The pyrites aren't gold, element 79 is.) So, the secondary intension of "there's gold in the mines" is false in this world, and the secondary intension of "gold has atomic number 79" is true there.

But when we consider that world as actual, we get different results. If our actual world turned out that way (with all the stuff in the mines around us being pyrite), then we'd say that gold turns out to be a pyrite, and that it turns out not to have atomic number 79. I.e., when we consider that world as actual, "there's gold in the mines" comes out true, and that "gold has atomic number 79" comes out false. I.e., the primary intension of "there's gold in the mines" is true in this world, and the primary intension of "gold has atomic number 79" is false there.

So for each of these statements, their primary intensions and secondary intensions give quite different results when evaluated in this world. This reflects the fact that the secondary intension picks out atomic number 79 in all worlds, but the primary intension certainly doesn't. Rather, the primary intension picks out something like the substance at the other end of a causal chain from the center. As far as this example is concerned we might be tempted by something like "the yellowish substance in the mines around us" as approximating the primary intension, but Kripke's illusion scenario suggests that that may need to be modified a bit (whether there is any modified description that can do the job is still controversial). In any case, despite the anti-descriptive intention of both the epistemic and modal arguments, we can see that their upshots for the primary and secondary intensions of terms like "gold" is quite different.

--Dave.

P.S. I hope this, and my other messages re the 2-D framework, help to flesh out the discussion in the book and the paper. You can consider these messages part of the "readings" for next week's discussion of 2-D issues. Sorry if I've belabored some of these points, but they're really crucial to what comes later, so I'm hoping all this will soon be second nature, if it's not already.

From owner-modality@LISTSERV.ARIZONA.EDU Sat Feb 13 23:52:03 1999
Date: Sat, 13 Feb 1999 23:10:50 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Proper Names
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Rachael writes:

>I have some questions about the Kripke reading- specifically on pages
>123-133.
> On page 123 Kripke acknowledges that we could be mistaken about gold
>having the atomic number 79. But he argues that given that gold does have
>the atomic number 79, something could not be gold without having the
>atomic number 79. Kripke makes light of Kant's mistaken assertion that
>gold is a yellow metal- but it seems that it is *possible* that just as we
>found out that all gold isn't yellow, we may find out that we are mistaken
>about gold having the atomic number 79. Kripke argues that we can pick
>gold out in all possible worlds by its atomic number. But if he was

>arguing 300 years ago, perhaps he might have made a similar argument- that
>gold is a yellow metal in all possible worlds.

OK, there are both epistemic and modal points here. I'll take the above sentence by sentence. Kripke's initial acknowledgment is the epistemic point that it isn't a priori that gold has atomic number 79. His following argument is the modal point that it's necessary that gold has atomic number 79 (given that it has that atomic number in the actual world). Kant's "mistake" is the epistemic claim that it's a priori that gold is a yellow metal. Kripke argues that this is not a priori, e.g. by raising the epistemic possibility of an optical illusion. You respond by in effect making the epistemic point that it isn't a priori that gold has atomic number 79, either. That seems right, and Kripke would agree. He thinks (modally) that gold's atomic number is necessary, but he doesn't think (epistemically) that it is a priori.

Note that Kripke's central point contra Kant isn't that we've found out that gold isn't a yellow metal (I don't think we've found this out!), but that we *might* find it out (witness the epistemic possibility of an optical illusion). Correspondingly, his modal claim that gold is necessary element 79 and not necessarily a yellow metal isn't based on the fact that gold has turned out to be element 79 and hasn't turned out to be a yellow metal -- as far as I know, it's turned out to be both. The point is that given the way things have turned out (in which gold is a yellow metal with atomic number 79, etc), it is necessary that gold has atomic number 79.

Your last sentence raises an interesting issue. 300 years ago, might Kripke have said that it was necessary that gold is a yellow metal? Well, presumably if he'd lived 300 years ago he'd have had very different philosophical views, but let's set that aside by thinking of old-Kripke as having developed a similar abstract philosophy, but as having different empirical knowledge (he thinks that gold is a yellow metal, but doesn't know about atomic number). Certainly the mere fact that he thought gold is a yellow metal wouldn't make him think its yellowness was necessary: as above, the inference isn't from "gold has property X" to "gold necessarily has property X". Rather, it's from "gold has deep property X" to "gold necessarily has property X", where the "deep" properties are some special class. Kripke-1970 thinks that gold's deep properties involve atomic number, not surface qualities.

The residual question, then, is whether Kripke-1670 might have thought that gold's deep properties involved its yellowness. I suppose he might have thought that gold was a homogeneous stuff with not many further properties than its surface properties. If he'd thought that, he might well have thought that it was necessarily yellow. If on the other hand, he thought it had some unified deep microstructure, maybe he'd have thought it had that microstructure essentially, even though he didn't know what that microstructure was. Things are complicated by the fact that the whole idea of microstructure as the unified basis of surface properties hadn't really developed then. So it's hard to know what to say. But in any case, it does seem that if Kripke-1670 had said something different about gold, it would have been grounded in some ignorance about the empirical facts.

Bringing the analogy back, Kripke-1970 probably should acknowledge that *if* he is wrong about the empirical facts about actual gold, then he could be wrong in his modal claims. For example, if it turns out that gold is a compound, then we might say it's that compound

necessarily, not element 79. Or if it turns out that gold has some even deeper structure, "quark structure XY", then maybe we'll end up saying that gold has quark structure XY necessarily, and perhaps not atomic number 79. But all that depends on the empirical facts, which Kripke is happy to take at face value from the scientists. If the empirical facts turn out differently, the specific modal claims may change, but the general philosophical claims will stay pretty similar (the inference from "gold has deep underlying structure X" to "gold necessarily has X").

Could even more radical empirical revision make us change even the abstract philosophical framework? Maybe that's not impossible, but it's not easy to see exactly how it would go.

> This brings up the question of what properties are essential. I
> think the answer is more difficult to pinpoint in cases like 'what makes a
> tiger a tiger' then for 'what makes gold, gold?' On page 121, Kripke
> seems to suggest that the external appearance of a tiger is not essential
> - we must look at its internal structure to determine whether it is a true
> tiger. My question is two-fold. First, if we were to discover a
> microscopic virus that happened to have the same internal structure as a
> tiger- would we recognize it as a tiger? Second, Kripke asserts that
> certain properties like quadrepedal, tawny yellow, and carnivorous are not
> essential to tigers, we might find that all these properties are optical
> illusions, but couldn't we discover the same of the internal structure of
> tigers?

The question of what properties are essential is a modal point. Kripke thinks that tigers have their DNA structure (or some such) essentially, i.e. that it is necessary that they have this structure. Your second point here brings together both modal and epistemic points. The claim that quadripedal/yellow/carnivorous aren't essential to tigers is a modal point: it's not necessary that tigers have these properties, since there are possible worlds where they don't. The idea that we might find out that these are optical illusions is an epistemic point: it's not a priori that tigers have these properties, since we might find out that they don't.

I think you're right that we could find out that tigers don't have the DNA structure, but that's an epistemic point (it's not a priori that tigers have this structure) rather than a modal point, so it doesn't threaten Kripke's claim that tigers have the DNA structure essentially (given that they have it in the actual world).

What if a microscopic virus turns out to have the same DNA structure as a tiger? That's an interesting one. Of course, as Larry points out, we're unlikely to find this out, and if we did discover it, it would cause our whole biological theory to be revised. My guess is that if we discovered this, we'd probably give up on the claim that DNA structure is the relevant deep structure, and so we'd give up on the claim that tigers have their DNA structure essentially: we'd have to find some other deep property that's responsible for the macroscopic differences. Whatever happened, it seems unlikely to say that we'd say that the virus is a tiger (though just possibly, there's some outlandish scenario where we would say they were members of the same natural kind, like carbon and diamonds?).

There's also the question, what should we say about a possible world where there is something with the DNA structure of a tiger but looks and behaves like a virus. That's a purely modal issue now (we're assuming in the actual world, the DNA and tiger-ish properties go together). Presumably there are weird possible worlds like this. Do

we call them worlds where tigers look like viruses, or not? My guess is that we wouldn't, and that we'd say the virus-ish things aren't tigers. If that's so, that means the property of being a tiger isn't identical to the property of having the DNA structure in question. It may still be that the DNA structure is necessary for being a tiger across possible worlds, but it isn't sufficient: perhaps one needs some sort of organic properties as well. E.g., one might argue that tigers are essentially animals, or essentially mammals, or some such, as well as being essentially that DNA structure (something can have more than one essentially property).

This raises more general questions about just what the essential properties of natural kinds are, and it's hard to give a complete answer easily. Personally, I find there are many cases about which I don't have firm intuitions, and it may be that beyond a certain point the facts about which properties are essential become indeterminate, or terminological, or some such.

> I guess I have a problem with how to identify natural kinds, >particularly in recognizing which properties are essential and which are >not. In light of past mistakes (like gold is a yellow metal) I am not sure >that we can ever know that we have actually discovered the essence of a >natural kind, or what picks it out in all possible worlds. Perhaps you >could say, given **our** concept of gold, what picks it out for us is the >atomic number 79. But if it seems logically possible to be wrong about >gold always having the atomic number 79 in this world, then why couldn't >we be wrong about it in other worlds?

That's a good point. Kripke will agree that insofar as we can be wrong about what gold is made of in the actual world, we can also be wrong about what's gold in other possible worlds. All he really wants to commit to is the claim that **if** gold has atomic number 79 as its deep structure in the actual world, then it has atomic number 79 across all worlds (or something like that). Presumably, he does actually believe that gold has that atomic number, so he also believes that gold has that atomic number essentially, but he needn't claim that he is certain of either (and they certainly aren't a priori). What matters here is more the abstract philosophical structure, and the conditional connection between gold's structure in the actual world and its structure across possible worlds.

(One might well argue that it's not a priori that gold has atomic number 79 as its deep structure, and it's not a priori that gold necessarily has atomic number 79, but it **is** a priori that if gold has atomic number 79 as its deep structure in the actual world, then gold necessarily has atomic number 79. At least, Kripke's philosophical methodology seems to suggest that he regards that conditional connection between nonmodal facts about the actual world and modal facts about possible worlds as a priori.

This is an instance of a general pattern that one might discern in Kripke's claims about a posteriori necessity. The necessary modal truths in question may not be a priori, but they are entailed a priori by certain nonmodal facts about the actual world. E.g., it isn't a priori that Hesperus is Phosphorus or that Hesperus is necessarily Phosphorus, but it is a priori that if Hesperus is Phosphorus (in the actual world), then Hesperus is necessarily Phosphorus. Similarly for other cases.)

> As our science develops we discover that our original theories >about light and heat were mistaken. I'm not sure how this works- but it

>seems like if enough of the data (what we recognize as heat and light)
>don't match up with the theory- then we find a new one. It is still
> possible to discover that heat is not just the motion of molecules- I
>suppose that to accomplish such a discovery we would have to be using our
>primitive notion of heat by which we identify it as a characteristic
>effect produced on our nerve endings. If heat just *is* motion of
>molecules then it looks like instead of reformulating the theory- we would
>have to call that other kind of thing that produces a hot sensation in us-
>something other than heat. It just doesn't seem that "heat is motions of
>molecules" or "water is H2O" are as necessary as "a bachelor is an
>unmarried man." What do you think?
> -Rachael

Here again, I think Kripke would agree that it's not a priori that heat is the motion of molecules, though it is necessary that heat is the motion of molecules (given that the motion of molecules causes heat sensations in the actual world). So we could find out that heat isn't the motion of molecules. If the scenario you describe turned out to be actual, I think Kripke would agree that we'd reformulate our theory, and say that the new thing that really causes the sensations is heat.

Kripke would insist, I think that "water is H2O" and "heat is the motions of molecules" are as necessary as "a bachelor is an unmarried man". He'd say that the difference is just that the latter necessity is a priori, while the first two are a posteriori.

I think there is a way of accommodating your idea that the first two are "less necessary" than the third, though. The third is both necessary and a priori. So in this case, the sentence has a necessary primary intension and a necessary secondary intension (it's true across all epistemic possibilities and across all counterfactual possibilities). Or in the terms of the book, it is both 1-necessary and 2-necessary. The first two, on the other hand, are necessary but a posteriori. So each of them has a necessary secondary intension but a contingent primary intension (they're not true across all epistemic possibilities). That's to say, they are 2-necessary but 1-contingent. (Or in the terms Evans uses, they are "superficially necessary" but "deeply contingent".) In effect, Kripke always identifies "necessity" with 2-necessity (necessity of secondary intension), so all these come out "necessary" by his lights. Still, the 2-D framework seems to isolate a reasonable sense in which the first two are "less necessary" than the third.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Feb 14 10:07:15 1999

Date: Sun, 14 Feb 1999 11:06:23 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Timothy J Bayne <bayne@U.ARIZONA.EDU>

Subject: Re: Proper Names

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Regarding the following, see below:

> Incidentally I heard someone a while back give a paper with the
> thesis, contra received wisdom but in sympathy with your view, that
> all uses of name tokens such as "Jane" are in fact tokens of the same
> name. So there is just one name "Jane", one name "Paul", one name
> "David Hilbert", etc. She addressed the problem of how the name gets

> to refer differently in different cases by arguing that names are
> implicit demonstratives -- in effect one is saying "that Jane", "that
> David Hilbert", etc, with different intentions to demonstrate in each
> case. A demonstrative like "that" can have different referents in
> different contexts and with different intentions, but it's still one
> word. Same on this view for "Jane". I'm not sure the view is
> completely unproblematic, but it's interesting.
>

Burge seems to have advanced something like this view. According to a secondary source, Burge argued in a 1973 JP piece that proper names don't refer to unique individuals, but to a kind of individual, namely the kind with individuals of that name in its extension. The idea, I take it, is that "Alfred" is a kind name that refers to all the individuals called Alfred. Does this mean that *Alfred* changes its meaning when another person gets called Alfred? I.e., Alfred did mean [x, y] when they only things called Alfred were x and y, but now it means [x, y, and z]? That can't be right. Perhaps the theory is that Alfred refers to the property (universal) of being such that Alfred is one's name - in which case (on most theories of universals), *Alfred's* meaning would not depend on the objects it was true of. *Alfred* would be a bit like *chair*, on this account. Being an Alfred would be a rather extreme nominalistic property.

Burge also brings in a demonstrative element to single out which Alfred one is talking about in a particular situation.

Anyway, apparently Burge used the fact that such sentences as 'there are lots of Alfreds in this department' to support his claim. Not having read Burge's paper I don't have a settled opinion of his account.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 19:07:58 1999
Date: Wed, 10 Feb 1999 19:07:54 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Do names have primary intensions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on the "minutes".

>Our discussion seemed to focus on two issues, broadly conceived.
>(1) What are primary intensions? We spent a lot of time trying to
>work out whether proper names, such as London, have primary
>intensions, and if so, >> what are they. There was some discussion of
>whether 'London' and 'Londres' have different primary intensions, if
>they have primary intensions at all.

Actually, I discuss the 'London'/'Londres' case a bit in "The Components of Content". Everyone should read that paper, since it gives a lot more detail on the 2-D framework, though it is primarily cast in terms of concepts, not words.

Do proper names have primary intensions? As I said yesterday, they arguably don't have PIs that are universal to all users of a name. It may be that two different users have different PIs for "London", just as for "Bill Clinton", etc. So this might lead one to say that names don't have PIs as a matter of "semantics", where the semantics of a term is supposed to be universal to all users of the term.

Still, I think that it is plausible at least that any name has a PI for any user on any occasion of use. We might put this by saying that "name tokens" have PIs (or are associated with PIs), even though "name types" may not. If so, we can arguably see PIs at least as part of the semantics of an "utterance" (cf. Reimer's distinction), and certainly as part of the semantics of an underlying thought.

Why think that name tokens have primary intensions? Basically, because like any referring word, a name (token) must have some pattern of application across epistemic possibilities. Just say I use the name "Godel". Then I can consider all sorts of epistemic possibilities about the way the actual world might be, and I can see how the name will refer with each of them.

Think of Kripke's own methodology in Lecture 2 -- we consider e.g. the epistemic possibility that the incompleteness of arithmetic was proved by a guy called "Schmidt" and stolen by a guy called "Godel" who published it, moved to Princeton, etc. Kripke notes, in effect, that if that epistemic possibility is actual, our term "Godel" refers to the Princeton guy. That's to say that the primary intension of "Godel" picks out the Princeton guy in the relevant centered world. One can do the same thing for a very wide range of worlds considered as actual (i.e. considered as epistemic possibilities): for a large numbers of such worlds, there seem to be clear facts about what our words will pick out if those worlds are actual.

We can think of the primary intension as the "reference-fixing" conditions of our terms, if we like, though actually I prefer to think of it as giving the epistemic application-conditions of our terms, or the "epistemic profile" (as opposed to the "modal profile" of the secondary intension). It's central to the way language works that any referring term has an epistemic profile (at least for a given user on

a given occasion): there are facts about how it applies to different epistemic possibilities, and about how it will refer if those possibilities turn out to be actual.

How does one evaluate the primary intensions of names that one uses? Basically, take a name, e.g. "London", take a given centered world, and ask oneself "to what does the name refer if that world is actual"? Here, one considers the world as actual, i.e. considers it as an epistemic possibility: "what if the world actually turns out to be that way?". And for a very wide range of such worlds, we have clear intuitions about how the name refers. E.g., if the actual world turns out to contain XYZ in the oceans etc, then we'll say that "water" refers to XYZ. If the actual world turns out to have the Princeton guy stealing the proof from Schmidt, we'll say that "Godel" refers to the Princeton guy. If the actual world turns out to have a non-whale-eaten guy at the other end of a causal chain from our use of "Jonah", we'll say that "Jonah" refers to that guy. Etc, etc.

Summing up a primary intension in language is often difficult, and the same goes for these cases; what really matters is the function from worlds to referents, not any capsule summary. But if I were to try to get at some of the things that are involved e.g. in the PI of my name "Godel", we might try: "the guy called 'Godel' who's at the other end of a causal chain from my use of the name", or something like that. That's imperfect, as Kripke argues -- e.g. one can consider epistemic possibilities in which the guy my term refers to wasn't called "Godel" at all (I've got his name wrong, or some such). But that just means that we have to refine our view of the PI. Importantly, Kripke's own methodology here relies precisely on evaluating how the term will refer if a given epistemic possibility turns out to be actual, i.e., on evaluating the PI of the term at a world.

As for "London" and "Londres": think of Pierre's situation. For him, there are lots of epistemic possibilities. The world he thinks he is in is one with a beautiful faraway city by the name of "Londres", and an ugly one close at hand called "London". Pierre quite reasonably says that if that epistemic possibility is actual (as he believes it is), then his terms "London" and "Londres" name two different cities: "London" picks out the faraway city and "Londres" the city close at hand. So in that centered world, the PIs of "London" and "Londres" give different results.

On the other hand, another epistemic possibility for Pierre (at least a broad epistemic possibility in the sense articulated earlier) is the actual possibility: i.e., that the people he got the term "London" from were actually referring to the same city he's living in now, and that it has both beautiful and ugly parts, etc. If Pierre were confronted with the hypothesis that this epistemic possibility is actual, he should rationally conclude that under that hypothesis, his terms "London" and "Londres" pick out the same city. So that's to say that in this centered world (Pierre's actual world!), the PIs of "London" and "Londres" give the same results.

So, Pierre's "London" PI and his "Londres" PI give the same result on some centered worlds (including the actual world), but different results on other worlds (e.g. the one he thinks is the case). So they are at least slightly different intensions overall.

How to summarize these intensions in language? Again, it's difficult and imperfect, but we might make a first attempt by saying that his "London" PI picks out "the beautiful city I've heard of under the name 'London'", and his "Londres" PI picks out "the ugly city I'm living

in", or perhaps "the city I've heard of under the name 'Londres'". Getting the details right will depend on careful consideration of cases, and will depend on just how Pierre's conceptual system is set up, etc. But this gives us enough to see how the PIs might pick out the same extension in one world (the actual world) but different extensions in some other worlds.

The fact that Pierre's PIs are different precisely reflects the fact that it is epistemically possible for him that London isn't Londres. This suggests a general principle: when "A=B" is a posteriori (for a user), A and B have different PIs. To see this, one can go through the following reasoning.

(1) "A=B" is a posteriori;

so (2) it is (broadly) epistemically possible that A is not B,

so (3) there is some epistemic possibility in which my terms "A" and "B" pick out different things (when that possibility is considered as actual)

so (4) there is some world in which the PI of "A" and the PI of "B" yield different extensions

so (5) A and B have different PIs.

I hope all the steps here make sense. You can illustrate it by thinking of the London/Londres case, or the Hesperus/Phosphorus case.

Note that the principle here is equivalent to: when "A=B" is a posteriori, "A=B" has a contingent primary intension. (N.B. the primary intension of a statement is just the obvious generalization of the primary intension of a term -- in the book I call this a "primary proposition", but I now tend to use the unified terminology.) If A and B have different PIs, there's a centered world where they pick out different extensions, and in which the PI of "A=B" is false. And to say that "A=B" has a contingent PI is just to say that there's a world in which the PI is false.

This is an instance of a very important general principle:

if a statement S is a posteriori, S has a contingent primary intension.

It's fair to say that this is at the heart of the two-dimensional account of a posteriori necessity. The central claim of this account is

if a statement S is necessary a posteriori, S has a contingent primary intension and a necessary secondary intension.

That's not too hard to illustrate by working through the standard Kripke cases. I note that there are just possibly some counterexamples to this principle which someone might put forward, though they probably won't be the standard Kripke cases. We'll be discussing potential counterexamples later in the course, but feel free to suggest any now.

But anyway, one can see how all this hooks in with the need for names to have PIs. Presumably a lot of statements involving names will be a posteriori for many or all users, e.g. "London = Londres", "Cicero = Tully", etc. That's just to say that for a given user, there is the epistemic possibility that the statement is false, i.e., there are some scenarios such that if they turn out to be actual, the statement

turns out false. So the very a posteriority of the statement indicates that the names in question have distinctive patterns of applications to epistemic possibilities, i.e. that they have (distinct) PIs.

That's enough for one message. Hopefully it gives useful background for thinking about issues in the rest of the minutes. More on the rest later.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 12 01:05:35 1999
Date: Fri, 12 Feb 1999 01:05:21 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Primary intensions and beliefs
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Brad writes:

>This question I think is partially addressed by Dave in his last
>posting, but here goes. I was wondering what constitutes a primary
>intension. In the examples we've discussed, such as "London" and
>"water", it seems that it is the beliefs that an agent has involving a
>particular concept which are constitutive of that concept's primary
>intension. So, for example, it is because Pierre believes that Londres
>is pretty and perhaps has various perceptual beliefs from photos (I
>think those kind of details were in the thought experiment?) that the
>primary intension of "Londres" differs from that of "London" (about
>which Pierre has very different beliefs, such as "it is ugly" and "it is
>dirty" and "the food there is awful").

Actually, I wouldn't say that beliefs "constitute" a primary intension, though there is a relationship between certain core beliefs and the primary intension. To determine what a PI is, one doesn't look at beliefs, but at the application-conditions of a concept to the world. What matters is what the concept picks out in any given epistemic possibility. So in Pierre's case, we look at what "London" would pick out for him if various epistemic possibilities turned out to be actual. That will be the primary intension.

Nevertheless, there is some sort of relationship between an agent's beliefs involving a concept and the concept's primary intension. We can think of a primary intension is roughly corresponding to the agent's "a priori beliefs" or "unrevisable beliefs" involving the concept. For example, if a user's primary intension for "one meter" picks out the length of a certain stick, this corresponds to the fact that it is a priori for that user that the stick is one meter long. This might be seen as a "core belief" that corresponds to the primary intension.

Note that not all of an agent's beliefs involving a concept will correspond so closely to a primary intension. Indeed, most beliefs won't be reflected in the PI at all. Most beliefs about meters, say -- that Olympic athletes run 100 meters, that there are about 1600 meters in a mile, etc -- and aren't reflected in the PI structure. This reflects that the beliefs in question are a posteriori, and could easily be given up if an appropriate epistemic possibility turns out to be actual.

In Pierre's case, just which of his "London" and "Londres" beliefs will be "core" and which will be "incidental" to his PIs depends to some extent on his psychology. I actually suspect that his belief that Londres is pretty isn't reflected in the "Londres" PI at all. It could epistemically turn out that the place he had heard of in Paris is actually ugly, but he'd still say that "Londres" refers to that city. If so, the PI (the application-conditions of the term) doesn't require prettiness. Perhaps a core belief might be something more like "Londres is the city I heard of a while ago under the name 'Londres'", or something like that; it's harder to see an epistemic possibility in which he'd give that up. As for "London", I suspect that ugliness and dirtiness will likewise be incidental. Maybe "London is where I've been living recently" is a core belief (i.e. he picks out the referent precisely as the place he's been living), or maybe "London is the place people call 'London'" is, depending on just how Pierre's psychology works.

In any case, I think it's a mistake to think about a PI too much in terms of beliefs. It's true that one can try to find core beliefs in the vicinity of a PI, but the fundamental thing is the pattern of application across epistemic possibilities. The beliefs that turn out to be "core" are just those that are seen to hold across any of these epistemic possibilities; but I'd argue that it's this pattern of application that tells us that the beliefs are "core", rather than vice versa. Otherwise, one can be confused by surface conditions of "coreness" that don't really reflect the deep epistemic profile of the concept. So beliefs are really secondary here.

>There is an article by Georges Rey (1983, I think) in which he argues
>roughly that we must distinguish the meaning of a concept from the
>beliefs that an agent has *about* that concept. I gather that on the 2d
>view Dave is giving, that distinction or separation isn't quite right.
>Some beliefs (and this is made clear in Dave's last posting) are
>constitutive of a concept's primary intension. And from Dave's last
>posting, it seems to be the beliefs that the agent him/herself takes
>(perhaps implicitly) to be constitutive.

Well, yes and no, as I think the above makes clear. A few things to say. (1) "Beliefs about the concept" taken literally are fairly irrelevant -- e.g., my belief that I have the concept "London". It's beliefs *involving* (or *using*) the concept that are at issue here -- e.g., my belief that London is pretty. I take it that's what you and Rey mean. (2) I agree with Rey that what really matters is the meaning (the PI), not the beliefs, although there may be a correspondence between the meaning and certain beliefs. (3) As above, even at the level of correspondence rather than constitution, it's only some beliefs that are relevant (the core ones).

All that being said, I think you're on the right track with what you say about the beliefs that matter being the ones that the agent implicitly takes to be constitutive. But again, I wouldn't want to *define* a PI in those terms.

>I think that, in the end, this seems ok to me, though I had previously
>been compelled by Rey's point. Perhaps Rey's point is applicable only
>to secondary intensions, which seem to be what Rey means by "meaning".
>There is the odd possibility that a person (say a chemist) comes to be
>so entirely convinced that water is H2O that the belief that water is
>H2O becomes constitutive of the concept "water". If future advances in
>science reveal that water is really XYZ (suppose that it turns out that
>the theory of the elements is entirely mistaken), on the 2d analysis my

>hypothetical chemist would have to conclude that water does not exist.

Hmm. I suspect that you're probably right that Rey's intended "meaning" is more like a SI. I think the best way to think about the chemist is not as someone who becomes really convinced that water is H₂O. Even such a chemist (normally understood) would presumably concede that she was wrong if the stuff out there turned out to be XYZ. That would suggest that her PI isn't really so H₂O-ish. Better to think of her as a theoretical chemist who has never had acquaintance with the stuff out there, who only knows molecular chemistry, and who uses "water" more or less as another name for H₂O. Then if the stuff in the lakes turns out to be XYZ, this chemist won't care at all -- she might go on using "water" as a name for H₂O (assuming H₂O is still taken to exist on the new theory). For her, "water is H₂O" really will be a priori.

[Technical note: The issue here is complicated by semantic deference, as in Burge's arthritis cases. If the chemist is like many speakers in intending to defer to usage in her surrounding community, then I guess it couldn't be that "water is H₂O" is a priori for her, and on making the XYZ discovery, strictly speaking her utterance of "water is H₂O" would be false. But let's say the chemist doesn't much care about the way others use the term, and uses it her own way (with her own full mastery of her own concept). Then the PI will work the way I suggested.

The residual issue (in the non-deferring case) is whether she could truly be said to have our word "water", or a different word altogether (one which happens to have the same sound and the same referent in the actual world). Some might argue that her failure to semantically defer, combined with her different usage, makes it a different word with a different meaning. Others might argue that the commonality in actual reference is enough, given that it's a natural kind term. They could arguably gain support from the way names work (it's epistemically possible that my "Paderewski" and your "Paderewski" could turn out to refer to different people, but as long as they actually refer to the same person, we're arguably using the same name). What to say about natural kind terms here is really a very subtle and difficult question. Fortunately it doesn't matter too much for our purposes.]

>Someone following Rey, however, would not have to say this. Instead, >one could claim that the chemist's *beliefs* about water are entirely >distinct from what his concept of "water" actually means. So it doesn't >matter how strongly the chemist believes that water is H₂O--if water is >really XYZ, the chemist simply has a false belief *about* water.

Well, I'll agree with Rey that the mere strength of belief isn't enough, but I guess I do think that some beliefs matter somewhere. In the first chemist case, we probably would say the chemist has a false belief about water, because it would turn out that her PI lies elsewhere. Similarly for the deferential theoretical chemist. But in the crucial case, the nondeferential theoretical chemist, if the XYZ scenario turned out to be actual, we might say that she doesn't have a false belief about water at all, just a true belief about something that isn't water (though she calls it "water").

(Note that in all three of these cases, there's arguably some match between PI and belief. Arguably, the PI of the first case would involve "watery stuff" and of the second case might involve "what people in my community call 'water'". These plausibly correspond to *other* beliefs of the chemist, e.g. the belief that water is watery

stuff or that it's called 'water' around here, which would then turn out to be the core beliefs. Of course in the third case the core belief would be that water is H₂O, and the PI would pick out the H₂O-ish thing in all worlds.)

>I certainly don't think that the above worry is a serious problem for
>the 2d account. It seems to me that the above might simply show that we
>find it difficult to imagine (or rather, that it is unlikely) that the
>primary intension of a person's concept of water would in fact come to
>be constituted by a belief that water is H₂O. What beliefs *are*
>constitutive of primary intensions, in general? That is, is there some
>general account to be given of the difference between constitutive and
>"incidental" beliefs (beyond the fact that the difference is that an
>agent *takes* certain beliefs to be constitutive and others to be
>incidental)? Perhaps the beliefs that are involved in the descriptions
>used for "reference fixing"?

The latter is more or less what I'd say. They can be seen as those involved in reference-fixing conditions, and might equally be seen as those that the agent would hold to be true no matter which broad epistemic possibility turned out to be actual. But again, the order of priority is not: give an account of what it is to be a core belief, and then define a PI in terms of that. Rather, its: define a PI in terms of application to epistemic possibilities (i.e. to worlds considered as actual), and then define core beliefs in terms of that.

>On a related note, I wonder if anyone can explain the following footnote
>from "The Components of Content" (my page 6, but I assume that varies):
>

>"But note that even Kripke's (1980) argument against description
>theories presupposes something like a primary intension. It proceeds by
>considering the referent of a concept such as "Godel" at various
>actual-world candidates, thus evaluating the primary intension, and
>arguing that any given description gives the wrong result."
>

>I'm not entirely clear how this case shows that Kripke presupposes
>primary intensions. Doesn't the "right result" for the reference of
>"Godel", on Kripke's account, turn out to be that the referent has to be
>*that man*--Godel! If I'm understanding Kripke correctly here, I can't
>see how this uses something like primary intensions rather than simply
>using something like a secondary intension--the referent in the actual
>world.

Hmm, remember that by this point, Kripke has moved away from the "modal" considerations in Lecture 1, and is now on to the "epistemic" considerations in Lecture 2. His point with "Godel" isn't to make the Aristotle/Hesperus/etc point that it picks out the same guy in all counterfactual worlds (that's the point about secondary intensions). His point is rather one about descriptions in the fixation of reference, and about epistemic possibility (so it's a point about primary intensions). In particular, he is taking a given description D (in this case "the person who proved the incompleteness of arithmetic" or some such), and he is arguing that it is not a priori that Godel is D. To do this, he considers an epistemic possibility (the one involving the proof being stolen) and suggests that if that possibility is actual, "Godel" picks out the stealer, not the prover.

So he is in effect making the point that the primary intension (the epistemic application-conditions) of "Godel" are not given by D. And similarly for any other description, he claims. Note that his methodology presupposes something like a primary intension, in

considering epistemic possibilities and evaluating reference in them. His claim is in effect that that the PI isn't given by a description, but rather works by returning whatever lies at the end of an appropriate causal chain from the word in question. (Of course Kripke himself wouldn't put things in those terms.)

Note that this epistemic point, that it is not a priori that Godel is D, is very different from the modal point, that it is not necessary that Godel is D (i.e., that the secondary intension of "Godel" isn't given by D). The latter point Kripke certainly accepts, and corresponds to the sort of point he makes in Lecture 1, but by now he is making a different point. (See the beginning of lecture 2 for explicitness on this.) I take it that your point about "Godel" referring to *that man* across possible worlds is an instance of modal point, not the epistemic point. (In effect, it's the claim that "Godel" is a rigid designator.)

The two points arguably come apart for a name like "Jack the Ripper", for which the modal point holds but the epistemic point arguably doesn't. Arguably, reference is fixed as "the person who committed the murders", and it's a priori that Jack the Ripper committed the murders (if he exists). Note that Kripke will still say that "Jack the Ripper" picks out *that very man* across possible worlds, so the modal point will still hold, even though in this case the epistemic point doesn't. Here, the primary intension of "Jack the Ripper" can be captured by a description, even though the secondary intension can't.

Translated into the 2-D framework, the modal point is that for all names, the secondary intension doesn't correspond to any description (except trivial ones). The epistemic point is that for most names, the primary intension doesn't correspond to any description. As I noted earlier, the modal point is less controversial than the epistemic point.

The distinction between the modal point and the epistemic point is subtle at first, but it's quite crucial to a full understanding of what Kripke's up to and of the 2-D framework, so people should fire away with questions if they're not quite clear on it.

--Dave.

P.S. Sorry about the unintended e-mail earlier.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 10 23:16:00 1999
X-Accept-Language: en
Date: Thu, 11 Feb 1999 00:09:35 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Brad Thompson <bradt@U.ARIZONA.EDU>
Subject: Primary intensions and beliefs
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

This question I think is partially addressed by Dave in his last posting, but here goes. I was wondering what constitutes a primary intension. In the examples we've discussed, such as "London" and "water", it seems that it is the beliefs that an agent has involving a particular concept which are constitutive of that concept's primary intension. So, for example, it is because Pierre believes that Londres is pretty and perhaps has various perceptual beliefs from photos (I think those kind of details were in the thought experiment?) that the

primary intension of "Londres" differs from that of "London" (about which Pierre has very different beliefs, such as "it is ugly" and "it is dirty" and "the food there is awful").

There is an article by Georges Rey (1983, I think) in which he argues roughly that we must distinguish the meaning of a concept from the beliefs that an agent has *about* that concept. I gather that on the 2d view Dave is giving, that distinction or separation isn't quite right. Some beliefs (and this is made clear in Dave's last posting) are constitutive of a concept's primary intension. And from Dave's last posting, it seems to be the beliefs that the agent him/herself takes (perhaps implicitly) to be constitutive.

I think that, in the end, this seems ok to me, though I had previously been compelled by Rey's point. Perhaps Rey's point is applicable only to secondary intensions, which seem to be what Rey means by "meaning". There is the odd possibility that a person (say a chemist) comes to be so entirely convinced that water is H₂O that the belief that water is H₂O becomes constitutive of the concept "water". If future advances in science reveal that water is really XYZ (suppose that it turns out that the theory of the elements is entirely mistaken), on the 2d analysis my hypothetical chemist would have to conclude that water does not exist. Someone following Rey, however, would not have to say this. Instead, one could claim that the chemist's *beliefs* about water are entirely distinct from what his concept of "water" actually means. So it doesn't matter how strongly the chemist believes that water is H₂O--if water is really XYZ, the chemist simply has a false belief *about* water.

I certainly don't think that the above worry is a serious problem for the 2d account. It seems to me that the above might simply show that we find it difficult to imagine (or rather, that it is unlikely) that the primary intension of a person's concept of water would in fact come to be constituted by a belief that water is H₂O. What beliefs *are* constitutive of primary intensions, in general? That is, is there some general account to be given of the difference between constitutive and "incidental" beliefs (beyond the fact that the difference is that an agent *takes* certain beliefs to be constitutive and others to be incidental)? Perhaps the beliefs that are involved in the descriptions used for "reference fixing"?

On a related note, I wonder if anyone can explain the following footnote from "The Components of Content" (my page 6, but I assume that varies):

"But note that even Kripke's (1980) argument against description theories presupposes something like a primary intension. It proceeds by considering the referent of a concept such as "Godel" at various actual-world candidates, thus evaluating the primary intension, and arguing that any given description gives the wrong result."

I'm not entirely clear how this case shows that Kripke presupposes primary intensions. Doesn't the "right result" for the reference of "Godel", on Kripke's account, turn out to be that the referent has to be *that man*--Godel! If I'm understanding Kripke correctly here, I can't see how this uses something like primary intensions rather than simply using something like a secondary intension--the referent in the actual world.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Feb 14 22:48:15 1999
Date: Sun, 14 Feb 1999 23:47:22 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>

Subject: Re: Epistemic and modal issues
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Hello, I know we're not directly discussing PI's in relation to reference anymore, but lately I think I've managed to unravel what little understanding I (thought I had) on this issue. So, apologies for re-opening some discussion, but here goes. A lot of what I took Thony's concern with PI's to be centered on was the convergence of PI's to descriptive content when using them to fix reference. I take it that Kripke (obviously) wants to avoid this possibility, requiring only of PI's that they are part of a causal chain that begins in the referent and links up to the subject (Chalmers too--see p.59 of TCM). My problem is that, pulling PI's away from descriptive accounts and towards causal ones (or at least calling them "compatible" with such accounts) does not make the application of PI's in reference fixing any less mysterious. I'm not sure that we can have cognitive access to things like causal chains, and if so, I'm not sure how any of this would work.

To put it simply, I don't see how something like "watery stuff" can fail to be descriptive (and therefore adopt all the problems of descriptive theories or reference), or (assuming it is not) how calling in something like a "causal chain" makes the application of a PI any clearer. If this is any objection at all then I suppose it would be sufficiently general to make reference fixing for SI's puzzling too. My general problem is that I can't seem to make sense of what we mean when we say "watery stuff" if it is not shorthand for some description of the properties of that stuff. If we want to call it a causal connection (involving some initial dubbing ceremony), then I guess I don't know how we have access to those sorts of connections such that we can accomplish reference fixing.

The point of this is just to express my confusion with exactly what is meant when someone begins talking about a primary intension. "Watery stuff" is clear enough, I suppose, but when I think about what I mean by this, I end up thinking of a bunch of descriptive conditions. There are two ways to run this, I guess. First, the confusion applies to PI's in particular, not SI's. This seems strange, since it would seem very arbitrary that "surface" properties but not the internal structure of something would be subject to description-theory objections (i.e., Kripke's). Second, fixing reference is equally problematic for PI's and SI's. This doesn't seem right either, because we want to preserve the notion that concepts like "watery stuff" apply in all possible worlds (to water) while "Water is H₂O" is different in that it could have been something else (XYZ) say, before the reference was fixed. So I suppose the modal distinction is preserved or insulated from epistemic worries about fixing reference, but I'm still troubled by the use and applications of things like PI's in the discussion so far, and I'm not sure what the point of the modal issues are if we can't make sense of PI's and SI's epistemically. Is this all clear (or has it been clarified already?). If so, disregard...

Erik L.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Feb 15 01:08:49 1999
Date: Mon, 15 Feb 1999 01:08:25 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Epistemic and modal issues
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Erik asks some important questions:

>Hello, I know we're not directly discussing PI's in relation to reference
>anymore, but lately I think I've managed to unravel what little
>understanding I (thought I had) on this issue. So, apologies for
>re-opening some discussion, but here goes. A lot of what I took Thony's
>concern with PI's to be centered on was the convergence of PI's to
>descriptive content when using them to fix reference. I take it that
>Kripke (obviously) wants to avoid this possibility, requiring only of PI's
>that they are part of a causal chain that begins in the referent and links
>up to the subject (Chalmers too--see p.59 of TCM). My problem is that,
>pulling PI's away from descriptive accounts and towards causal ones (or at
>least calling them "compatible" with such accounts) does not make the
>application of PI's in reference fixing any less mysterious. I'm not sure
>that we can have cognitive access to things like causal chains, and if so,
>I'm not sure how any of this would work.

Actually, this posting is right on topic for this week's discussion of
PIs and the 2-D framework.

Are PIs like descriptions? On my view, in a sense yes, and in a sense
no. PIs needn't be linguistic formulae, or correspond to such
formulae. A PI is a function from centered worlds to extensions.
Sometimes such a function will be completely captured by a
description, such that the PI picks out whatever satisfies the
description in world: e.g. "the length of such-and-such stick", for
the case of 'meter', or "the individual at the center of the world",
in the case of 'I'. But this sort of neat expressibility may be quite
rare, and certainly isn't universal. The conditions of application in
a PI may be too unruly to be captured by any simple formula, or
perhaps by any finite formula at all. So we shouldn't expect to
always be able to express PIs as simple descriptions.

But still, there's a looser sense in which PIs behave something like a
description. Like a description, a PI captures the conditions that an
object in the actual world has to satisfy in order to be the referent
of one's term or concept. And like a description, it's plausible that
a PI is cognitively accessible in principle, at least in the sense
that if a given centered world is specified, a thinker can in
principle figure out the value of the PI in that world (by figuring
out what the extension of our concept or term would be if that world
were actual).

Furthermore, one can at least attempt to *approximate* the conditions
involved in a PI by a description. Any short description may be
imperfect -- there may be cases where the description and the PI give
the wrong results, as e.g. in the Kripke cases. But arguably one can
give better and better approximations to the PI by giving longer and
more involved descriptions, to handle exceptions and the like. It's
not clear that any finite description will be perfect: maybe it will
be like the case of "knowledge", where one seems to get longer and
longer descriptions (cf. Chisholm's thirteen clauses, and Shope's The
Analysis of Knowing), and one gets the sense that one needs some sort
of infinite "fractal" structure to get all the details just right.
But in any case one can arguably give better and better descriptions
that cover more and more of the cases.

How does all this play out, if something like the causal theory of
reference is correct? Well, I wouldn't say that the PI in this case
is a causal chain: you're right that we don't have access to causal
chains. But note that even in the Kripke cases, we still have

conditional access to extension. Kripke's very own methodology is to consider various scenarios as epistemic possibilities, and to figure out what the extension of our terms (e.g. "Godel", "Feynman", "gold") would be if those scenarios turn out to be actual. So we still seem to have cognitive access from a function to from scenarios to extensions. And that's a primary intension.

How can one cash out the primary intension in descriptive form here? Well, as usual, it's not obvious that one can. But one can at least provide approximations. E.g., one might approximate the PI of 'Godel' as "the individual at the other end of a causal chain from my use of the term 'Godel'", or "the person that experts in my community call 'Godel'", or "the person I've heard of under the name 'Godel'", or some such. Some of these are vague (e.g. the first -- what sort of causal chain is required?) and some may be imperfect (e.g. the second -- maybe I misheard and evryone else calls him "Girdle", not "Godel"?), but at least they give a loose approximation. And arguably they can be refined into better and better approximation.

Is there a perfect description? Well, that's still a topic of some debate. Some think that something like the third description above can do the job (metalinguistic descriptivism). And some think that if the causal theory of reference is right, one ought to be able to bundle it all into a description, something like "whatever individual is causally connected to me in the manner specified by the causal theory of reference" (causal descriptivism). Even Kripke thinks that something like the latter might be done (see p. 162 and footnote 38), though he thinks it is trivial. Whether a less trivial description can do the job -- maybe a "cluster" description with causal and metalinguistic elements -- is still arguable.

Anyway, the PI framework doesn't need to take a stand on whether the PI can be captured by a description. I think by now you can see in what sense it is "description-like", in that it in effect gives the conditions a referent must satisfy, and is cognitively accessible. And in any given case, one can at least try approximating the PI by a description. I often give descriptions to express something about a PI myself, but one shouldn't forget that the description is usually just an approximation. But in any case, given that we have conditional access to extension in a given case (as Kripke's own methodology suggests in even his anti-description cases, and as I think is plausible in almost any case), it follows that we have, in effect, cognitive access to a PI.

>To put it simply, I don't see how something like "watery stuff" can fail
>to be descriptive (and therefore adopt all the problems of descriptive
>theories or reference), or (assuming it is not) how calling in something
>like a "causal chain" makes the application of a PI any clearer. If this
>is any objection at all then I suppose it would be sufficiently general to
>make reference fixing for SI's puzzling too. My general problem is that I
>can't seem to make sense of what we mean when we say "watery stuff" if it
>is not shorthand for some description of the properties of that stuff.
>If we want to call it a causal connection (involving some initial dubbing
>ceremony), then I guess I don't know how we have access to those sorts of
>connections such that we can accomplish reference fixing.

Does the above help? Again, "watery stuff" is really a term of art, and isn't really meant as a perfect descriptive characterization of the PI. Using "watery" in its intuitive sense, then the description might give a very loose approximation to the PI, but there are some cases where it might come apart. In the book i think I stipulate

instead that "watery stuff" is a term of art to capture the conditions involved in the PI of 'water', whatever they are. To give a full accounting of those conditions and of the PI, one can really only consider things case by case: look at scenarios, and see how the extension of 'water' should come out there.

If I were to really try giving a description for the PI of "water", I suppose it would probably involve its superficial properties (though defeasibly, to allow for the epistemic possibility of optical illusion?), and it might also involve a causal connection criterion, perhaps even a condition of causal connection to my use of the token 'water'. And any such description would likely be imperfect. But at least it gives an intuitive idea.

In any case, the complexities of exactly capturing the PIs of our linguistic terms don't really matter much for our purposes in metaphysics. Beyond a certain point, the complexities of "what counts as 'water' in this situation, or in this one" become somewhat terminological. For explanation, for example, what matters is explaining the manifest properties of people such as Godel and stuff such as water, and here the terminological complexities of just who and what gets to count as "Godel" or "water" becomes irrelevant after a certain point. (Even if the PI of 'water' isn't perfectly captured by "watery stuff", it might as well be for many purposes.) It remains the case that getting things just right is very important in the philosophy of language, but it's not clear that all the fine details matter to the metaphysician or to the scientist. That's why I tended to brush aside some of the fine details in the book.

>The point of this is just to express my confusion with exactly what is
>meant when someone begins talking about a primary intension. "Watery
>stuff" is clear enough, I suppose, but when I think about what I mean by
>this, I end up thinking of a bunch of descriptive conditions. There are
>two ways to run this, I guess. First, the confusion applies to PI's in
>particular, not SI's. This seems strange, since it would seem very
>arbitrary that "surface" properties but not the internal structure, of
>something would be subject to description-theory objections (i.e.,
>Kripke's).

Re SIs, I suppose things are a bit different since it's not part of the package that they are cognitively accessible (in a nontrivial sense) to anyone who possesses the concept. You need substantial empirical knowledge of the actual world in order to evaluate the SI of one's concept in an arbitrary possible world (e.g., you need to know about the actual structure of water to know that the SI of 'water' picks out H₂O across worlds). But I suppose there's a sense in which at least after the fact, SIs are often easier to summarize with descriptions (e.g. "H₂O") than PIs are. I'm not sure that this is always so (personally, I think one might even worry about whether any H₂O in any world counts as water), and it's not clear that there aren't some more resources for expressing PIs (I'll talk about alternative ways of looking at them a bit later), but even if so, I think it's really an issue of relative simplicity, rather than expressing some deep metaphysical difference. (Of course, there is the epistemological difference in cognitive access.)

>Second, fixing reference is equally problematic for PI's and
>SI's. This doesn't seem right either, because we want to preserve the
>notion that concepts like "watery stuff" apply in all possible worlds (to
>water) while "water is H₂O" is different in that it could have been
>something else (XYZ) say, before the reference was fixed. So I suppose the

>modal distinction is preserved or insulated from epistemic worries about
>fixing reference, but I'm still troubled by the use and applications of
>things like PI's in the discussion so far, and I'm not sure what the point
>of the modal issues are if we can't make sense of PI's and SI's
>epistemically. Is this all clear (or has it been clarified already?). If
>so, disregard...

Well, we want "water is watery stuff" to be true across all worlds considered as actual (i.e., PI true in all worlds), and "water is H2O" to be true across all worlds considered as counterfactual (i.e., SI true in all worlds). Of course one difference is that the necessity of the SI depends on empirical facts about the actual world while the necessity of the PI doesn't. But that just reflects the fact that PIs are much more closely tied to epistemic notions such as apriority than SIs are.

Re epistemic access to PIs and SIs: as I said, I think one has access to a concept's PI merely by possessing the concept (you then have the ability to figure out how it refers if a given world is actual). For an SI, I think one only has conditional access to it: one doesn't have nontrivial access to it merely by possessing the concept, but one does have access given sufficient empirical knowledge. So on my view, one has a priori access to a PI, and a posteriori access to an SI, but one does have some sort of a priori access to the way that an SI depends on the way the actual world turns out.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Feb 15 22:02:48 1999

Date: Mon, 15 Feb 1999 23:01:21 -0700

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Angela J Burnette <aburnett@U.ARIZONA.EDU>

Subject: Re: Epistemic and modal issues

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

all

re Dave's answer to Erik's comments about PI's, the questions that follow have really already been answered but I guess I'm just not getting it through my thick skull...

Dave says that primary intentions are functions from centered worlds to extensions, which is straightforward, but given that it may be impossible to specify or spell out the conditions for application, I (still, sorry) don't see how such a function can do us any good...(a function that has no specifiable conditions of application hardly seems like a function at all)

I understand that the PI is supposed to "capture the conditions that an object in the real world must satisfy in order to be the referent of one's term or concept", but given that this must be done on a case by case basis and cannot be cashed out in terms of any description (even if one is available, such a description is not of primary importance, right?)...I still don't understand the "sense" in which anyone who has the concept can "in principle" evaluate PI's except as a matter of personal intuition...it seems to me like everyone will have their own "function" for each PI, which will also be just what they happen to have in mind...I think that Dave intends this not to be a problem....something about not mattering "after a certain point"...this is what I really need explained...why doesn't it matter after a certain point?

I understand what a PI is intended to accomplish and I understand how, in principle, or in some vague sense, it is actually used, but what I'm unclear on is why it is "unimportant" that PI's be such vague and unspecified tools...

angela

From owner-modality@LISTSERV.ARIZONA.EDU Tue Feb 16 06:26:56 1999
Date: Tue, 16 Feb 1999 06:25:01 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Epistemic and modal issues
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Angela asks some good questions, to which I'm not sure I have perfect answers, but I'll do my best.

>Dave says that primary intentions are functions from centered worlds to extensions, which is straightforward, but given that it may be impossible to specify or spell out the conditions for application, I (still, sorry) don't see how such a function can do us any good...(a function that has no specifiable conditions of application hardly seems like a function at all)

Hmm, I'd say a function is a function (i.e. a mapping from one domain to another, here a mapping from worlds to extensions), irrespective of whether there is any easy summary of a "rule" that generates the function. You're right that's its nice to have some sort of summary of a function when one can, but it's not clear to me why it should be essential. In the case of PIs, we've seen that one can get reasonably good approximations of PIs by descriptions, and better and better approximations by more and more complex descriptions. I'd like to think that that's good enough (and in fact, that even the approximability of a PI by descriptions is something of a bonus rather than being essential. But maybe I'm wrong. What do you think is the reason or purpose for which exact descriptive summaries of PIs are essential?

>I understand that the PI is supposed to "capture the conditions that an object in the real world must satisfy in order to be the referent of one's term or concept", but given that this must be done on a case by case basis and cannot be cashed out in terms of any description (even if one is available, such a description is not of primary importance, right?)...I still don't understand the "sense" in which anyone who has the concept can "in principle" evaluate PI's except as a matter of personal intuition...it seems to me like everyone will have their own "function" for each PI, which will also be just what they happen to have in mind...

Right (re your first question). Sometimes there will be a description, but even here we figure out the aptness of a description precisely by seeing whether the evaluation of a description matches our evaluation of the PI at a given world. (Witness Kripke's methodology: he argues against various descriptions by arguing that there are worlds where the description gives the wrong referent. That's to say, the evaluation of the description doesn't match our evaluation of the PI. So there is arguably some sense in which even when a description is available, the PI is prior to the description, at least epistemically. (Though it may be that at some other level, e.g. psychologically or even semantically, it's the description that fixes the PI.)

You're right that "intuition" plays a central role in evaluating PIs. Again witness Kripke's methodology in the theory of reference: he considers his own intuitions about what the referent of "Godel", "Feynman", etc, will be across various situations. One might say this makes it just "personal intuition", but note that we do seem to find considerable agreement across many or most such cases (cf. the proof-stealing scenario), so it's not as if mere personal taste is at play. It seems to be a fact about us that once we know about how the actual world turns out qualitatively, we're in a position to know the reference of our terms and the truth-value of our statements (we'll be focusing directly on this "scrutability" thesis later on). This doesn't appear to be a mere vagary of personal taste, but a reasonably systematic process of reasoning and judgment.

I note that if it's just "personal taste" across possible cases, by similar reasoning it would be just personal taste that determines our judgments of reference and truth-value in the *actual* world (given empirical information about it). On the face of it, that doesn't seem too plausible: although intuition is involved in these judgments (is in almost any judgment), there's plausibly a fact of the matter in both cases about the reference of our concepts and the truth-value of our thoughts, a fact of the matter that we are plausibly capable of figuring out given sufficient qualitative empirical information and a priori reasoning. But if we can do that for qualitative information about the actual world, presumably we can also do it for information about all sorts of other ways the actual world might hypothetically turn out. Just apply the same sort of reference- and truth-figuring processes (after all, if we were all told that the actual world was just that way, we'd be able to figure out the consequences for truth and reference). So I do think something more than just personal taste is present here.

Of course it may be that two users will make slightly different judgments about the truth and reference of their terms across various possible worlds (considered as actual). One way this can happen is for there to be conceptual differences between the two -- i.e. our concepts really do have slightly different PIs as an aspect of their content, as e.g. in a deferential user of "arthritis" and an expert user, or two people with slightly different concepts of "London". But that isn't a problem. Leaving that sort of case aside, I suppose another thing might happen is that we'll make slightly different judgments in borderline or fuzzy cases. I think that sort of thing is relatively unimportant. When it comes to borderline and indeterminate cases, all bets are off. What matters is that we have clear judgments (in principle) about the clear cases.

I also note that it's not clear how having descriptions would help with this problem. If two users have different judgments about reference in a given case, presumably that would require that they have different descriptions. So if you're right that there are a lot of differences in judgment, then there will be a lot of personal descriptions, too. So the issue of intuition and judgment is arguably somewhat independent of the issue of intensions vs. descriptions.

[Side note: There is a question about whether a PI should be defined in terms of a user's best a priori *judgments* about the reference and truth-value of their concepts and thoughts in given epistemic possibilities, or should be defined as what the reference and truth-value of the concepts and thoughts *would be* in those epistemic possibilities, if they turned out to be actual. I hold that these two

notions always or almost always give the same results, but for some purposes it is useful to distinguish them: call them an epistemic PI and a fixing PI.

Re your point above, a user's judgments and intuitions play a role in defining an epistemic PI (though I note that one is idealizing by allowing arbitrary a priori reasoning here). For fixing PIs (which are arguably the most straightforward), however, judgments and intuitions play no role in the definition. All that matters there is patterns of reference and truth. So focus on those if you want to avoid any problems with intuition and judgment.

We can talk more about epistemic PIs and fixing PIs if people like. The issue will come up a fair bit anyway, re the discussion of scrutability. Basically, epistemic PIs as defined are a priori accessible, and fixing PIs as defined fix reference, but one needs epistemic PI = fixing PI in order to have something a priori accessible that fixes reference. The requisite bridge is provided by the "scrutability" thesis that reference of our concepts and truth-value of our thoughts are always knowable in principle given sufficient qualitative information about the world and sufficient a priori reasoning. That's a thesis I've already implicitly appealed to a few times, and which is consistent with e.g. the standard methodology in the theory of reference.

Some might think there are exceptions to this thesis, "inscrutable truths": e.g. if the epistemic theory of vagueness is true, it might provide some. If this were so, it would raise complexities for PIs, e.g. requiring us to distinguish epistemic and fixing PIs. I hold that there are no inscrutable truths, but that's something we'll be discussing at length later.]

>...I think that
>Dave intends this not to be a problem....something about not mattering
>"after a certain point"...this is what I really need explained...why
>doesn't it matter after a certain point?

I don't remember what I said didn't matter. I think small disagreement about the application of our concepts across cases. That basically suggests that our concepts have slightly different PIs, or that one of us hasn't really reasoned things through, or that we're dealing with a borderline case. But it's not clear that any of those phenomena matter much. Different people do indeed have different conceptual contents (often even associated with the same word), and a little indeterminacy around the edges of a PI doesn't matter any more in principle than indeterminacy around the edges of actual-world application. A general theme is that whatever goes for the actual world goes for possible worlds, too (e.g. re indeterminacy, epistemic access, and the like).

>I understand what a PI is intended to accomplish and I understand how, in
>principle, or in some vague sense, it is actually used, but what I'm
>unclear on is why it is "unimportant" that PI's be such vague and
>unspecified tools...

Well, it depends on what you mean by vague and unspecified. "Vagueness" in the sense of indeterminacy at the edges (i.e. worlds in which a PI is borderline true/false) can happen but doesn't seem worse than actual-world vagueness. "Vagueness" in the sense of indeterminacy even on core cases I don't think exists -- for any concept and thought, there is a very wide range of scenarios on which the concept is true or false. "Unspecified" in the sense of

ill-defined I'd also argue with -- PIs have a perfectly good definition (whether epistemic or fixing), associating a concept with a clear value in worlds across a range of cases.

Your central problem is presumably not any of the above, but rather "unspecified" in the sense of not being given by a neat description. I'm still not sure just why this is a problem. I guess the proof has to be in the pudding, but I think that application to a range of issues shows that the notion of a PI can do a lot of explanatory work (see e.g. the content paper), and that indeed much or most of the work that descriptions do for us is in virtue of their determining an intension rather than in virtue of their linguistic form.

There's more to say on this, but I have to go to bed. I may well be missing something, and I imagine you're not the only one who is at least somewhat bothered by the issue, so you or anyone else should fire away and say just what you think we lose by having an intension instead of a description.

--Dave.

P.S. I'm looking forward to everyone else's posts for this week, and to the minutes of Tuesday's meeting. Note that re the "Components of Content" paper, the issues about psychological explanation and belief ascription, etc, aren't central to what we're focusing on this week, though we may come back to them. It's the general framework that's most important, and its application to the understanding of necessity, apriority, reference, and content.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 00:27:27 1999

Date: Wed, 17 Feb 1999 01:26:31 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik A Herman <erikh@U.ARIZONA.EDU>

Subject: 2D

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

First off, I am in agreement with the idea of notional content. I have always considered the thinkers relation to the concept to be an integral part.

I've been wondering if/how time fits into the picture. It is mentioned in a few spots in Components of Content:, "facts about the current environment...", "Tom is hungry at time t..." Take the Phosphorus/Hesperous example, if I understand correctly, these have different primary intension but fix the same referent, Venus. Can I say that there is a physical difference, namely the space and time, between Phosphorous and Hesperous?

Also, I don't understand what happens in the case where someone ONLY knows about H2O's chemical properties (theoretical chemist). Is this a case of a descriptive concept where the secondary intension is simply the primary intension?

My intuitions about Josh's brain in a vat, as we discussed this afternoon, are that, similar to a paradigm shift in science, water would just mean watery stuff like always, but now it just has a different explanation that the "new man" would need to learn to assign it.

Erik H.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 02:42:22 1999
Date: Wed, 17 Feb 1999 02:42:10 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: 2D
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Erik writes:

>I've been wondering if/how time fits into the picture. It is mentioned in
>a few spots in Components of Content:, "facts about the current
>environment...", "Tom is hungry at time t..." Take the
>Phosphorus/Hesperous example, if I understand correctly, these have
>different primary intension but fix the same referent, Venus. Can I say
>that there is a physical difference, namely the space and time, between
>Phosphorous and Hesperous?

Well, time is built into the center of the world. In effect, the center consists of an individual at a specific time (or an ordered pair of an individual and a time). That's needed for concepts such as "now", which pick out the current time (i.e. the time at the center), and for a bunch of related concepts.

I'm not sure exactly what you're asking in the last sentence. Basically, the idea is that the PI of "Hesperus" picks out (something like) the bright object visible in the evening sky from the point of view of the center, and the PI of "Phosphorus" picks out the bright object visible in the evening sky from the point of view of the center. In the actual world, the evening object is the morning object, so those two intensions pick out the same object, namely Venus. So the two concepts have different PIs but the same referent (cf. Frege on sense and reference).

>Also, I don't understand what happens in the case where someone ONLY knows
>about H2O's chemical properties (theoretical chemist). Is this a case of
>a descriptive concept where the secondary intension is simply the primary
>intension?

I'd say that's close to right. At a first approximation, we can say the PI and the SI both pick out H2O in all worlds, so it's like a descriptive concept. On a second approximation, though, one can note that "hydrogen" probably isn't a descriptive concept. Its PI picks out some gaslike stuff in all worlds, and its SI picks out some particular atomic structure in all worlds. Similar for "oxygen". So it may be that the SI of "H2O" requires just that atomic structure (with just the right number of protons, neutrons, electrons, etc), while the PI of "H2O" isn't quite so constrained.

(The PI of H2O is tricky, though. It comes down to tricky questions about e.g. what we'd say if hydrogen and oxygen exist but are quite unlike what we thought they were in atomic structure; but nevertheless there is still a substance with two hydrogen particles and one atomic particle, or some such. Would we count that as H2O, or would the change in chemical theory make us throw away the whole concept of "H2O" and say it doesn't refer? Maybe the former, but it's a subtle issue. Maybe in the grey area. In any case, though, it does seem that the PI will allow for at least some smallish variation in structure that the SI arguably won't.)

Interesting to see people's comments re Josh's case. It would be good if people can be as explicit as possible about how their views on the referent of "water", etc, in these cases fit into the 2-D framework. (E.g. given your views on the reference of the concept and the truth of the thoughts before and after, what is the PI, and what does it pick out in the various centered worlds before and after vat-release?)

A little more on Angela's note from yesterday. Maybe one can think of a PI as something like a tacit descriptive content of a concept. It may not be something that a speaker can directly articulate, and there may not even be any summary of the relevant content in words, but one could argue that in some sense the descriptive content is there at least tacitly, in guiding one's judgments about reference in various epistemically possible cases. Of course we may have to break the link between "descriptive content" and precise linguistic descriptions, though. It will be more like something analogous to descriptive content with the linguistic element removed. (Something like a property that determines reference, or some such.) That's a bit loose, but maybe it helps with the intuitive picture.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Feb 16 21:59:38 1999
Date: Tue, 16 Feb 1999 21:59:21 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Josh raises a couple of interesting brain-in-vat scenarios. There are actually two sorts of question in the vicinity: (a) what to say about the beliefs in question, whether they are true or false before and after escaping from the vat, and (b) whether what one says re (a) can be accommodated within the 2-D framework, and if so, how (what do we have to say about the PIs of the relevant concepts, etc). Personally I'm not certain of the right thing to say about (a) (there are multiple intuitions), but I'd like to think that any reasonable answer to (a) can be accommodated within the 2-D framework somehow. It's a nice test case, though.

Maybe I'll hold off on posting my own response to these scenarios in order to give other people a chance to post their analyses and comments on what's going on here. Fire away!

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Feb 16 13:53:15 1999
Date: Tue, 16 Feb 1999 14:50:32 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Sorry to be out of the loop the last week. Its possible that in catching back up I've missed something. If something like the following has already been discussed I apologize.

The following is a case that I don't know what to do with. Suppose scientists pull your brain out of the vat it has been in up until now and tell you that your whole life has been a lie. They created the world you thought you were living in all this time. Suppose you are convinced they are right. But now you are out of the vat and in a real body and can go explore the world. As you leave the laboratory and walk out into the world you see that everything is just like it was in your vat. There is grass, trees, water etc. But then one of the scientists runs up and says, "Ahh. One point we forgot to tell you. The underlying physics of the world is different from the one we presented to you in the vat. From the molecular level on down everything in your vat world was different. For example, water is XYZ."

Case 2a: Same as above except that while I'm still living in my vat world, a dramatic scientific revolution occurs in which physics, from the molecular level on down is determined to be different than we thought. For example, water is XYZ. I'm shown how thinking that water is H₂O is wrong and actually it is XYZ. After I'm convinced of the new theory I'm pulled out of the vat and everything runs as above except that now the real physics matches my vat physics.

Now here is the puzzle. Primary intentions pick out "watery stuff" in the actual world. My intuition is that the actual world in both cases is the post-vat world. However, if I were the one pulled out of the vat in case 1 I'd say that there is no 'water' in the actual world. There is just watery stuff. But if I were pulled out of the vat in case 2, I'd say there is water in the actual world.

If you share these intuitions, then there seem to be two possibilities:

1) My primary intention of water includes the chemical make up of water or 2) Primary intentions do not fix reference in the actual world but fix reference in some other way.

What do you all think?

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Tue Feb 16 23:56:08 1999

Date: Wed, 17 Feb 1999 00:55:16 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik J Larson <erikl@U.ARIZONA.EDU>

Subject: Re: 2D questions

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

This is a fascinating case, but I'm not so troubled by it primarily because I don't share the feeling that water in post-vat world would be relegated to "watery stuff" if the underlying physics of the world turned out different. In a sense, the scenario is not appreciably different than one in which my "brain in the vat" life was just general ignorance of the way the world really was. For the purposes of fixing reference with a PI, you have to buy that the shift in physics from vat-world to post-vat-world will shift your PI of water, but it's not at all obvious that this would happen. A more plausible scenario (for me, anyway) would have one replacing the physics of water while retaining the PI to fix reference for whatever was watery stuff, in the vat or actual world. Here I think epistemic issues pull away from the modal concerns. Now, given that the PI of water does not change from vat to actual world, the physics of water is not part of its PI, so the first possibility can perhaps be dispensed

with. The second possibility is sort of conditional on the first, in the sense that, if one doesn't think that the PI will shift from vat to actual worlds, then the reference fixing work of the PI is intact. Of course, things get substantially trickier if one shares Josh's intuition on the vat to actual world shift in PI (from water to non-water) from the physics-of-water epistemic shift in these worlds.

ErikL.

to be out of the loop the last week. Its possible that in
> catching back up I've missed something. If something like the
> following has already been discussed I apologize.
>
> The following is a case that I don't know what to do with. Suppose
> scientists pull your brain out of the vat it has been in up until now
> and tell you that your whole life has been a lie. They created the
> world you thought you were living in all this time. Suppose you are
> convinced they are right. But now you are out of the vat and in a real
> body and can go explore the world. As you leave the laboratory and
> walk out into the world you see that everything is just like it was in
> your vat. There is grass, trees, water etc. But then one of the
> scientists runs up and says, "Ahh. One point we forgot to tell you.
> The underlying physics of the world is different from the one we
> presented to you in the vat. From the molecular level on down
> everything in your vat world was different. For example, water is
> XYZ."
>
> Case 2a: Same as above except that while I'm still living in my vat
> world, a dramatic scientific revolution occurs in which physics, from
> the molecular level on down is determined to be different than we
> thought. For example, water is XYZ. I'm shown how thinking that water
> is H₂O is wrong and actually it is XYZ. After I'm convinced of the
> new theory I'm pulled out of the vat and everything runs as above
> except that now the real physics matches my vat physics.
>
> Now here is the puzzle. Primary intentions pick out "watery stuff" in
> the actual world. My intuition is that the actual world in both cases
> is the post-vat world. However, if I were the one pulled out of the
> vat in case 1 I'd say that there is no 'water' in the actual world.
> There is just watery stuff. But if I were pulled out of the vat in
> case 2, I'd say there is water in the actual world.
>
> If you share these intuitions, then there seem to be two possibilities:
> 1) My primary intention of water includes the chemical make up of
> water or 2) Primary intentions do not fix reference in the actual
> world but fix reference in some other way.
>
> What do you all think?
>
> Josh
>

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 00:17:25 1999
Date: Wed, 17 Feb 1999 01:16:34 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Angela J Burnette <aburnett@U.ARIZONA.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

re: Josh's scenario...

I guess I just don't share the intuition that I, upon exiting the vat and discovering that water was XYZ, would from that point on be of the opinion that there really was no water. Rather, I would just think that I had been deceived, or had been radically mmistaken about what water is....

I understand the temptation, especially given the amount of scientific information that is part of common knowledge, to say that H2O really is part of the primary intension of water, but, the very fact that I (and presumably others) can make sense of counterfactuals like, "Water could have been other than H2O," or "Water might not have been drinkable by humans", or "Water might not have had the property of appearing blue in daylight," etc...strongly suggests that almost nothing of what we know a posteriori of water is contained within the primary intension.

best, Angela

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 00:34:32 1999
Date: Wed, 17 Feb 1999 01:32:28 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik A Herman <erikh@U.ARIZONA.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

In response to Josh's scenario: my intuition is that the two cases are identical-- and that the new watery stuff is "water", it's just that he/we were wrong about it's makeup.

Erik H.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 12:58:07 1999
Date: Wed, 17 Feb 1999 13:52:38 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Regarding the responses to my BIV cases, I'd like to make a few clarifications. There are two ways that I see of accounting for my intuitions. The first is to claim that the chemical make up of water is part of my PI. The second is to claim that my PI is just like everyone elses BUT PIs don't map centered worlds considered as actual onto referents.

I think the second of these claims is the more likely. The idea would

be that PI's perhaps map centered worlds thought to be actual during concept formation onto referents. (I already see several problems with this definition, but that is the basic idea.) Since vat-Josh formed his concepts while thinking that vat-world was actual, he would then map water to clear drinkable liquid in vat-world. And in Vat-world water is H₂O.

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 18:10:24 1999
Date: Wed, 17 Feb 1999 18:07:56 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Here are a few thoughts on Josh's "water" case. In this message I'll give my own analysis, and in the next I'll give some thoughts on others' analyses.

The first thing to get straight on is whether one thinks the judgments in question are true or false, before and after the vat.

Case 1: I think "water is H₂O" in the vat. I escape from vat, find a similar-seeming world, except that the watery stuff is XYZ.

Case 2: I think "water is H₂O" in vat, until a within-vat revolution convinces me of "water is XYZ". I escape from vat, find a similar-seeming world, and the watery stuff there is XYZ.

Q1: What (if anything) do I refer to as "water" pre-escape?
Q2: What (if anything) do I refer to as "water" post-escape?
Q3: Is my statement "water exists" true or false pre-escape?
Q4: Is my statement "water exists" true or false post-escape?

There are various things to say here. It's probably simplest to first set aside the belief changes re XYZ and H₂O, and think about what "water" might refer to pre- and post-vat in an ordinary BIV case. Here some distinct issues peculiar to BIV cases come up. For example, many would say that while in the vat, my term "water" doesn't really refer to anything, and that my thought "water exists" is false. After all, isn't a brain in a vat a classic case of someone who is deceived about the external world?

One could accommodate this line in the 2-D framework by saying e.g.: (1) the PI of "water" picks out something like "the watery stuff in the environment that the being at the center has been causally related to" (not an implausible analysis of my "water" PI), and (2) if I am a BIV, my centered world (i.e. my actual world, not merely the world-as-I-believe-it-to-be) doesn't have anything that satisfies that PI. (There's no watery stuff causally related to me in my actual environment.) So the PI will pick out nothing, and my vatted thought "water exists" will be false.

On this line, what will happen when I come out of the vat? Well, arguably when I emerge for the first time into the "real world", my term "water" doesn't yet refer to actual water, because there's no causal connection yet. But after interacting with the environment for a while, the causal connection will get going, and perhaps eventually my term "water" will come to refer to the actual stuff (H₂O). Note

that all that will be compatible with the PI analysis above: nothing will satisfy the PI at stage 1 (centered on me in vat) at stage 2 (centered on me just after escape), but something may satisfy it at stage 3 (centered on me well after escape), as by then the center will have plenty of causal relations with some watery stuff.

Another line that some philosophers take is that while in the vat, my terms like "water" refers to some chemicals or nutrients in the vat around the brain, because these are what I'm causally connected to. I don't find this very plausible myself, but if one takes this line, one could accommodate it by dropping "watery stuff" from the PI above, and by saying that the PI picks out roughly "the stuff in the environment that is causally responsible for the being at the center's use of the word 'water'". In this case, we'll have "water exists" true at stage 1 (it picks out chemicals), arguably false at stage 2 (it still picks out chemicals, but I suppose that could be true since the chemicals still exist), and true again at stage 3 (by now, the PI picks out real water).

Alternatively, one could be tempted by the idea that while in the vat, one's term "water" refers to something in one's "virtual world", or something like that, so that one's vatted thoughts "water exists" is true. This has the problem of making it hard to express how the BIV is deceived, and could even end up leading to some sort of idealism or phenomenalism. But there is arguably something to the intuition that there's some sense in which the BIV's "water is H₂O" is truer than "water is XYZ" (assuming the virtual world is an H₂O-world). Maybe one could accommodate this by saying there's two ways to read such thoughts, a strong way in which they all come out false, and a "weak" way where some come out true, or some such. E.g., the "weak" way might involve prefixing "According to the virtual world", or might have a PI involving "seeming-objects which seem to be causally connected to me", or some such. This would be a tricky row to hoe, but maybe there are possibilities. In this case, one could say that reference of "water" gradually shifts after release, just as above.

Anyway, I think one can factor away from these tricky BIV issues for considering Josh's main point. I'm not sure whether Josh's point would work equally well if the vat world were replaced by another planet in which one is brought up, and that brought to earth. That would probably give similar results at least to line 3 above, and arguably to line 2, though maybe not to line 1. But the crucial claims about how "water" refers after release may be similar each way.

What to say about case 1, where I believed "water is H₂O", and find out post-vat that the watery stuff is XYZ? I'd be tempted to say, as in the cases above, that immediately post-release my term "water" still refers to whatever it referred to just pre-release: either nothing, or some chemicals, or perhaps some virtual H₂O in the virtual world. So my statement "water exists" then will be false, or in any case if true won't be made true by the actual watery stuff. But after a while of interacting with the real environment, my term "water" might well come to refer to the XYZ, and my claim "water exists" will be made true by the actual environment.

I think that's so despite the fact that I believed in the vat that water was H₂O. After all, within the vat that belief was potentially malleable (if people had told me the watery stuff was XYZ, I'd have accepted that water was XYZ), so the same ought to be the case outside the vat. In fact, I think that even if no-one outside the vat ever tells me that water is XYZ, my term "water" may eventually come to

refer to the XYZ by virtue of my causally interacting with it, etc. In this case, I'll eventually have a false belief that "water is H₂O" but a true belief that "water exists".

If that's how things so, I think it's compatible with the PI staying constant throughout. E.g. if the PI is "the watery stuff in the environment that the being at the center has been causally related to", this seems to give the right results: it picks out nothing at stage 1, nothing at stage 2, XYZ at stage 3. And that's so irrespective of my belief changes.

I'd be tempted to say something similar about case 2. Here, I think there still needs to be a gradual reference-shift from vat-reference to actual reference. Even if one thought "water is XYZ" in the vat, and actual watery stuff is XYZ, it's not clear that one's thought (immediately post-escape) "water is XYZ" is true, or at least it's not clear that it's made true by the actual XYZ. It takes some interaction for the term "water" to pick out the actual stuff.

[A side note: It may be that if the scientists tell you immediately "you were in a vat, but now you're out, and fortunately the real world is a lot like you thought the vat world was", you might decide to make a new stipulation that all your words will refer to actual things rather than vat things, rather than keeping a lot of terms around that don't refer to anything useful. That's tricky in some cases: e.g. re "my wife", or "Ethel", say, does one really want to say that the term refers to a person in the actual world who you've never strictly speaking met before? But I suppose one could try to stipulate that at least some terms refer to their qualitative counterparts. If that's so, then "water" might immediately pick out the respective watery stuff.

As for what to say if there are a few small differences between vat world and actual world (e.g. they replaced H₂O by XYZ), that's awfully tricky. I guess it depends on one's stipulation. One might stipulate "in case of any differences, my term doesn't refer", or one might stipulate "in case of any differences, my term refers to whatever is superficially similar enough", or some such. In the first case, my post-escape "water exists" might at first be false; in the second; it might be true. But that all depends on just what stipulation one makes. Such a stipulation might well change one's term's PI a little bit, as any new stipulation will.

Probably it's easiest to think about these things in a case where the scientists don't tell you you've escaped, so one isn't faced with the immediate wrenching discovery that forces you to reevaluate your whole conceptual scheme. Instead, maybe do things so that you don't know you've undergone such a radical shift. Even so, we can evaluate truth and reference, and we can still consider what happens if there turn out to be small changes in the environment. That way the "stipulation" aspect is removed.]

Anyway, that's how I'd be tempted to analyze these cases. I think it's plausible on most ways of doing things that one's PI stays constant, but its referent may gradually shift, in effect due to a change in the centered world in question (or at least, a change in the location of the center). The one case where my PI may change is the case where I know that I've shifted environments, and decide to restipulate the use of my terms; but that's a very special sort of change.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 17:24:53 1999
Date: Wed, 17 Feb 1999 18:24:00 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on Josh's BIV case and on de se representation. . .

The BIV case is a nice case, and a difficult one. I had an answer, but talked myself out of it. Here's another stab. This is in three parts, all rather convoluted.

(1) Part of what is up for issue is whether the primary intension of our concepts like <water> are egocentrically indexical, or whether they are more communal. Suppose that Josh is the only BIV, and that he thinks of water as <the stuff that causes (most of) most of *my* watery experiences>. On this construal of the 1-intension, Josh's water concept refers to certain types of electrodes, or computer programs, or whatever it is that causes most of his watery experiences. (Why do we describe it as his "water concept"? Cos it's connected to his watery experiences, which are connected to our water concept.) This, I take it, is the Putnamian position on BIVs. On a *communal* interpretation of the 1-intension of water, it is <the stuff that causes most of the watery experiences of most of the folks around here.> (There are lots of indexicals in it). Now, whether this concept still refers to the electrodes or programs of whatever depends on whether Josh is the only BIV around. If the vast majority of the folks around here are also BIVs, then perhaps one could argue that even now it still refers to these electrodes. (Of course, the problem is that Josh isn't communicating with any of these - presumably. So maybe the communal condition can't get a toe-hold in giving the referent of an individual BIV's concept.)

But let's suppose that the 1-intension of Josh's <water> is communal, and that he's the only BIV around here. Then, his thoughts about water would be mostly mistaken cos they are about water, and he isn't connected to water in the right way. Although of course his belief that water is H₂O would be true if the stuff that causes most people's watery experiences is H₂O. Of course, it may not always be easy to tell whether one of his thoughts is true or false, because it may not be clear whether it is about water. Suppose that he thinks that he is having a shower: 'Ah, a lovely shower, this stuff sure is hot' and such like. This might be a true thought if the stuff that is causing his experiences is indeed hot. But insofar as he thinks that the stuff that is hot is the same kind of stuff that everyone else (i.e. non-BIVs) thinks is hot when they are taking a shower, his thought is false. Similar comments applying to his thought, 'It sure is good to run naked through the water of this Bondi surf.' This is true, insofar as it is good to run naked through the Bondi surf, but of course it is false in that Josh-the-BIV is not running anywhere, least of all running naked through the Bondi surf.

(2) Let's push the Putnamian line for a minute and see where it goes. Take Josh-the-BIV's water concept to refer to whatever causes his watery experiences. Post-vat Josh's water concept also refers to the stuff that causes his watery experiences. These two concepts (mental representations) are similar in that they are both related to watery experiences, that is, they are both object concepts that are related to the *phenomenal concept* <watery>. But should we think that there is any sense in which VAT-Josh and post-Vat Josh have a concept, i.e. <water> that can be tracked from

the VAT to after the VAT? I wonder. Both in the Vat and after the vat Josh has a concept that is related to <watery>, but that doesn't seem sufficient to justify the claim that he has a single concept that survives the experience from VAT to actual world, that we can reidentify as the same concept. After all, both my concept of a certain chemical substance (say, glump - suppose that glup is watery) and my concept <water> are connected to <watery stuff>, but that doesn't give me any incentive to identify <glump> with <water>. In order to identify the concept that we are calling Josh's pre-Vat water concept with his post-Vat water concept it needs to be the case that most (much?) of what he has learnt about water in the VAT holds true in the actual world. Does envatted Josh's knowledge of the microstructure of water play an important role here? I don't see why. On an inferential role account of concepts, beliefs about the microstructure of a substance might be just one sort of inferential node among many. If the microstructure of XYZ is such that it behaves in almost all of the same ways that H2O behaves, then we might be inclined to say that Josh has learnt that water is XYZ, or whatever it is in the actual world. (Of course, XYZ can't behave in all of the same ways that H2O does, otherwise we would be able to discover that some stuff is H2O rather than XYZ.)

To sum: supposing the individualistic spin on water, in which the primary intension is <whatever has been causing *my* watery experiences> we (and Josh) might still be tempted to identify his in-Vat watery concept with his post-Vat watery concept, even though it refers to a different stuff, because many of the inferential connections that he learnt while in the vat still hold true in the actual world. If post-Vat Josh thinks of his envatted experiences as including water thoughts, then it is clear that he must have a water concept that we can trace from the Vat to the real world.

(3) Here's a similar case to Josh's that puts it in a communal setting. Suppose that we are all shifted to Twin Earth in our sleep. Twin Earth is as the travel brochures describe it: it's the same as earth, except the predominant watery stuff in the environment is XYZ. After a couple of days we would discover that "this stuff is XYZ". I guess most of us would say, "Huh, I guess that water isn't H2O after all, it's XYZ." But it seems that we would all be mistaken. Water is H2O, but this ain't water. Of course, after a couple of years on twin earth, 'water' would come to refer to XYZ. So it seems.

Well, I'm not sure what the point of this scenario is. Maybe it raises the question: 'Is the 1-intension of <water> "the stuff around here that causes most of our watery experiences", or is the 1-intension of water "the stuff that is causing and has caused most of the watery experiences around here at the moment and in the past." The suggestion is that the 1-intension of natural kind terms anchors them to the past (and maybe the future).

Tim

University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 20:31:28 1999
Date: Wed, 17 Feb 1999 20:29:40 -0800

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on analyses of Josh's case by Josh, Erik, Erik, Angela, and Tim. N.B. This presumes the previous message with my own analysis.

Josh wrote:

>Now here is the puzzle. Primary intentions pick out "watery stuff" in
>the actual world. My intuition is that the actual world in both cases
>is the post-vat world. However, if I were the one pulled out of the
>vat in case 1 I'd say that there is no 'water' in the actual world.
>There is just watery stuff. But if I were pulled out of the vat in
>case 2, I'd say there is water in the actual world.
>

Like others, I'm not sure I share your intuitions here. Maybe there's some reason for saying that post-escape, your term "water" doesn't refer to anything in the actual world. But I think one can argue when it does begin referring to something in the actual world, it will refer just as soon in each case. After all, the only difference between the cases is that in one I believe "water is H2O" and in the other I believe "water is XYZ". This might make a difference to reference if these were "core beliefs" for me and unrevisable, or some such. But in fact it seems that these beliefs are quite revisable, a posteriori beliefs, as reflected by the fact that upon "discovering" in the vat that watery stuff is XYZ, one would say "water is XYZ" is true, not that water doesn't exist. I think the same goes outside the vat. Insofar as my word refers to stuff outside the vat at all, it's in virtue of its role as watery stuff, and it doesn't matter whether it turns out to be XYZ or H2O.

>If you share these intuitions, then there seem to be two possibilities:
>1) My primary intention of water includes the chemical make up of
>water or 2) Primary intentions do not fix reference in the actual
>world but fix reference in some other way.

Hmm. I don't share the intuitions, but perhaps one can make sense of your intuitions by saying: the PI of water picks out "whatever has the same structure as the stuff I've been related to in my past environment", and going for a view (the third line in my previous post) on which in the vat environment, you're related to virtual-H2O and virtual-XYZ in case 1 and case 2 respectively. Then post-vat, out in the real XYZ-world, your "water exists in this environment" might come out true in case 2 but not in case 1 (because XYZ has the same structure as virtual-XYZ, but not as virtual-H2O). That would seem to give the results of your intuitions, though I don't know whether (a) it gets at what underlies your intuitions, or (b) it's plausible.

[N.B. If things work this way, it's immediately post-escape that the changes will show up, rather than later. I.e. even before causal interaction with the new stuff, my "water exists around here" will be true in case 2 rather than case 1. With enough causal interaction, reference might still shift so that the claim will be true in both cases.]

This analysis may get at something about what you're saying re "PIs

don't fix reference in the actual world". I'd say that even on this analysis, they do fix reference in the actual world, but the actual world is a 4-D spacetime manifold, i.e. it includes all the facts about the past. And here, your PI will fix reference partly by appealing to facts about your past environment (i.e., the way things are in the past environment of the being at the center of the world) rather than your current environment.

(It's plausible that something like that applies to many cases. E.g., if I am transported to Twin Earth without knowing, then presumably my immediate claim "there is water around here" will be false, even though there is watery stuff (in the loose sense) around there. Essentially, this is because the PI of water fixes reference to something like "the watery stuff I have causally interacted with", which involves reference to my past environment as well as my current environment. Of course if I live on Twin Earth long enough, my reference may shift enough so that "there is water around here" will be true; that's also predicted by this PI.)

Looking at your (Josh's) more recent second message, maybe something like this analysis fits.

>Regarding the responses to my BIV cases, I'd like to make a few
>clarifications. There are two ways that I see of accounting for my
>intuitions. The first is to claim that the chemical make up of water
>is part of my PI. The second is to claim that my PI is just like
>everyone else's BUT PIs don't map centered worlds considered as actual
>onto referents.

>
>I think the second of these claims is the more likely. The idea would
>be that PI's perhaps map centered worlds thought to be actual during
>concept formation onto referents. (I already see several problems
>with this definition, but that is the basic idea.) Since vat-Josh
>formed his concepts while thinking that vat-world was actual, he would
>then map water to clear drinkable liquid in vat-world. And in
>Vat-world water is H₂O.

Well, one problem with what you say above is that if I grow up (in the actual world) thinking water is XYZ, then this formula will predict that my claim "water exists" is false, which doesn't seem right. And more generally it seems that it will be hard to have false beliefs while one is growing up, which seems problematic. Maybe it's better to say that something special is going on with the vat case -- these aren't just any old false beliefs about the world, but a whole virtual environment, one that is so self-consistent and central that it deserves to count in some sense as my "actual environment" while I'm in the vat. If so, then one can argue that my vat terms fix reference to things in my virtual environment, and that even after leaving the vat, they still involve the way things were in the vat environment.

Note again that this doesn't involve ignoring the actual world, but it does involve concentrating on a particular aspect of it, i.e. my past environment, which in this case will be my past virtual environment (which of course is in some sense part of the actual world).

On this way of doing things, the escape from vat to world is just like going from Earth to Twin Earth (in case 1) or like going from Twin Earth 1 to Twin Earth 2 (in case 2). Either way, one shifts environment, and one's initial reference is fixed to the stuff in one's original environment, so the truth of one's immediate-post-move claim "there is water around here" depends on whether there is anything around here with the same structure as the stuff in one's

past environment. So true in case 2 but not in case 1. That seems to accommodate your intuitions, even with a constant PI that fixes reference via the actual world (including past environment).

Anyway, note that one this way of doing things, what makes the differences between the cases isn't just that one had different beliefs in the past about one's environment -- it's that in some sense, one actually had different environments in the two cases (albeit virtual environments, in Josh's version). To see the difference, think about what we'd say if (a) the virtual environment "really had" XYZ structure in it (i.e. if you'd looked under a virtual microscope, that's what you'd have found), but (b) you believe the whole time that there is H₂O there (you never look under a microscope). I predict that in this case, Josh will say (or ought to say) that "water is H₂O" is false when thought in the vat, and false after escape, while "water is XYZ" is true both times. That suggests that it's the past environment that matters, not the past beliefs.

Of course one can question (a) whether a "virtual environment" is really enough of an environment to make concepts refer and to make beliefs true, and even then one could question (b) whether the "real" XYZ environment really qualifies as having the "same sort of stuff" that is present in the virtual environment (one might argue that there are deep differences between virtual XYZ and real XYZ, so that the "virtual" concept won't pick out the "real" stuff after escape). But assuming one can make a case for the appropriate answer on those two questions, then Josh's position will come out as reasonable, and the PI above seems to capture roughly what's going on.

Re Erik L's, Angela's, and Erik H's messages: I think my intuitions largely agree with these, i.e. that there's no major difference between the two cases: insofar as case 2 "water" picks out the actual stuff post-vat, so does case 1 "water". (Although after going through the above, I have to say that I feel at least some of the tug of Josh's way of doing things, at least on the "virtual environment" reading.) I think one still needs to distinguish what happens immediately after release from what happens after a period of adjustment, though. I'm tempted to say that there is an element in the PI of "water" requiring a previous causal connection to the stuff, so it doesn't refer to the actual stuff immediately (in either case), but only after a while.

If one doesn't agree with the causal requirement, one could go with a looser "the watery stuff around here now" PI, so that I'll come out referring in both cases immediately after escape. But I think the causal requirement is independently plausible.

I do largely agree with what Angela says here:

>I understand the temptation, especially given the amount of scientific
>information that is part of common knowledge, to say that H₂O really is
>part of the primary intension of water, but, the very fact that I (and
>presumably others) can make sense of counterfactuals like, "Water could
>have been other than H₂O," or "Water might not have been drinkable by
>humans", or "Water might not have had the property of appearing blue in
>daylight," etc...strongly suggests that almost nothing of what we know a
>posteriori of water is contained within the primary intension.

One thing to note is that "could have" and "might have" counterfactuals are often ambiguous between a PI and a SI reading, i.e. between (broad) epistemic possibility and Kripkean "metaphysical"

or "subjunctive" possibility. I'm not 100% sure which (or both) you're meaning to appeal to here. Arguably one needs both to make best sense of these counterfactuals.

If one accepts Kripke's subjunctive intuition, "water could have been other than H₂O" will be true in the epistemic (PI) sense but not the subjunctive (SI) sense. As for "water might not have been blue", etc, this is most straightforward for me to make sense of on the subjunctive (SI) sense: one pictures the actual stuff, i.e. H₂O, having a different appearance. On the epistemic (PI) sense, this is harder to make sense of, since water's appearance properties are relatively close to its PI, i.e. are "relatively a priori" of the "water" concept. But presumably one can still make sense of them by considering certain strange (broad) epistemic possibilities, e.g. on which there has been an optical illusion, or water's surface properties are very different from how one thinks they actually are. That reflects the fact that the beliefs in question aren't so "core" as to be unrevisable. As one builds in more and more core surface properties of water, though, this gets harder to do, and the epistemic "might have" claim becomes harder to make sense of.

Re Tim's message:

(1) Tim raises a possibility I didn't mention in the last message, that my vatted "water" concept may be deferential, fixing reference via my community. Still, I'd argue that it seems wrong to say that I defer to the people out there in the real world outside the vat, and it also seems wrong to say that I defer to other BIV's who happen to be out there. (It doesn't seem right to say that those people's reference makes a difference to my reference.) If I'm deferring, it's to people in my "virtual" world, the ones I think I "got" the term from. (I suppose it could be that those virtual people in my world are "avatars" of "real" people or of other BIVs, so I'd be deferring indirectly to those people, but I'll set that possibility aside.)

If that's so, then I don't think deference makes an essential difference to the picture. If the non-deferential concept picks out nothing, or chemicals, or virtual XYZ, so will the deferential concept, on this way of looking at things. The main difference will be the precise nature of the "causal chain" connecting me to the referent, i.e. on whether it needs to go through other "virtual people", or not.

(2) Tim also raises the question of in what sense I'll have the same concept inside and outside the vat. This seems most straightforward in the case where I don't know I've shifted. In that case, there seems to be the same sort of continuity that one would find if I go from one world to another without knowing. I'll deploy a lot of concepts, and they certainly won't be a whole new repertoire that I've acquired overnight -- rather, they'll be the old concepts (although a lot of my resulting beliefs may now come out false). It may be that eventually reference will shift, so that in a certain sense I'll have different concepts -- I'll at least have concepts with a different referent and SI, though not a different PI -- but that's a slightly different phenomenon.

There's also the special case where I'm told I've shifted, and where I make a new stipulation about my concepts so my beliefs don't come out useless. That would plausibly involve a "different concept" in a reasonably strong sense, but that also seems to be something of a special case.

Tim raises some worries about concept identity:

>But should we think that there is any sense in which VAT-Josh
>and post-Vat Josh have a concept, i.e. <water> that can be tracked from
>the VAT to after the VAT? I wonder. Both in the Vat and after the vat Josh
>has a concept that is related to <watery>, but that doesn't seem
>sufficient to justify the claim that he has a single concept that survives
>the experience from VAT to actual world, that we can reidentify as the
>same concept. After all, both my concept of a certain chemical substance
>(say, glump - suppose that glup is watery) and my concept <water> are
>connected to <watery stuff>, but that doesn't give me any incentive to
>identify <glump> with <water>.

I presume you're talking about the vatted concept and the long-post-vat concept. In favor of "same concept" is that they have the same PI, and a continuous history of use between them. Against "same concept" is that they have a different referent and SI. What answer one gives will partly depend on whether one individuates concepts by PI, SI, referent, or history of use. I think this is somewhat terminological. I have some temptation toward "same concept", given the centrality of PIs, but then the difference in referent and SI is pretty significant, too. As for water/glump, I'm not sure how to imagine the case -- do they pick out the same referent, or not? If the same referent, presumably we have same SI, different PI, so the reverse of the above. If a different referent, presumably we have different PI and SI, which is worse (presumably even if the PI of "glump" involves "watery", it will be in a different way, as it isn't a priori that water is glump). Either way, one relevant difference is that the vat case involves sameness of PI and continuity of use, whereas this case doesn't.

>To sum: supposing the individualistic spin on water, in which the primary
>intension is <whatever has been causing *my* watery experiences> we (and
>Josh) might still be tempted to identify his in-Vat watery concept with
>his post-Vat watery concept, even though it refers to a different stuff,
>because many of the inferential connections that he learnt while in the
>vat still hold true in the actual world. If post-Vat Josh thinks of his
>envatted experiences as including water thoughts, then it is clear that he
>must have a water concept that we can trace from the Vat to the real
>world.

I think I agree with this, more or less. There are actually two issues: (1) sameness of concept pre- and post-vat, and (2) whether the concept picks out the actual stuff. That the "inferential connections still hold true in the actual world" seems to be getting at (2) rather than (1). In my terms, I'd put it by saying that as long as the actual world isn't *too* different from the vat world, actual watery stuff can come to satisfy the PI of my "water" concept (though possibly with a time lag). I'm inclined to think that (1) doesn't turn so much on the actual facts, but more on facts about the subject's psychological structure (modulo issues about individuating concepts by referent). That seems to fit with your last sentence above.

(3) Tim raises the case where I'm shifted to Twin Earth in my sleep. As I said in the last message, I think Tim is right that my "water" won't initially refer to XYZ, though it may after a couple of years. I think the moral of that is that the PI of "water" anchors one to one's past environment, as Tim suggests. Basically because it picks out (at least in part) the stuff that one has causally interacted with

in the past. In fact I think that for many of these cases, the present environment is mostly irrelevant (except to determining whether one's beliefs are now true or false, etc). The future environment even more so, though maybe one could come up with cases where it matters.

All thoughts and reactions are welcome!

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 18 16:44:56 1999
Date: Thu, 18 Feb 1999 12:47:46 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: Minutes from Tuesday
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Here are the minutes from Tuesday. I thank Brad for filling in some details.

The first thing we discussed was why we should care about the primary intentions. Roughly the line was, "if water *is* H₂O then why do we care what water could have been. It was suggested that the reason is that if, via the PI we could determine that it was logical impossible for water to just be H₂O, then we know that we have more to look for. This, of course, is the basic line for consciousness.

The bulk of the rest of the discussion was on proper names again. Thony argues that if the PI of a name is what picks out its referent, then two names which pick out the same object in our world could pick out different objects in another world. But it seems that as a matter of logic, if 'a=b', where 'a' and 'b' are names, then necessarily 'a=b'

I suggested that in this framework proper names are not being used as labels for objects. Rather they are functions which pick out objects of a certain description.

I think Thony and I now agree that the question is whether names ought to work this way.

> Brad tried to provide motivation for thinking that names do in fact have
> descriptive content. One such motivation comes from considering cases in
> which co-referring names can't be substituted salva veritate (such as in
> belief contexts). He also mentioned substitution failures in "simple
> sentences" (see Saul 1997), such as "Lois Lane kissed Superman before she
> kissed Clark Kent".

> We also briefly discussed again the matter of what the primary intensions
> of names typically are (such as in the London/Londres case discussed
> before).

We then discussed my BIV cases. From an intuition point of view we were evenly split about whether or not there was really any water after you were pulled out of the vat. Brad suggested that it may just be very difficult to think of this case as anything other than a person jumping between two possible worlds. This, of course, is impossible. But it does look kind of like that.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 12:01:52 1999
Date: Wed, 17 Feb 1999 12:59:13 -0700
Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Anthony T Lane <atlane@U.ARIZONA.EDU>

Subject: possibility?

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

I think this is probably getting somewhat off the topic, but I have a general worry about making sense of secondary intensions.

The 2-intension of water is H₂O, given that, in this world, we take it that this chemical structure explains the various properties we associate with water. More generally, it seems that secondary intensions are different from the primary intension only when there is some sort of reductive explanation of a particular term or thing (p62).

It is not, of course, directly observable that water is H₂O. Rather, it is the result of theory constructed from a variety of experimental results. Consider a possible world in which water is H₂O, but in which the natural laws are such that the elements behave radically different than they do in our world. Suppose, for example, that Hydrogen atoms form extremely unstable compounds when they are in the presence of other Hydrogen atoms. In this possible world, the dominant 'watery stuff' is XYZ. Thus, XYZ is the stuff picked out by the primary intension of water. XYZ behaves experimentally much like H₂O behaves in our world. But XYZ is not composed of two Hydrogen atoms and an Oxygen atom in this possible world.

It seems to me that the secondary intension of water cannot just be H₂O. It must also include a host of other information regarding the underlying physical theory that makes this identification meaningful. If there is a possible world in which XYZ behaves for the most part like H₂O, only X, Y and Z are not 2 H's and an O, it seems strange that there is any value in the secondary intension of water picking out H₂O, stuff that is, in this possible world, unstable.

If this isn't just confused, it may be that this is a reason why you say primary intensions are particularly important to your account.

Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 01:44:18 1999

Date: Fri, 19 Feb 1999 01:44:04 -0800

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: Minutes from Tuesday

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

A few comments on the minutes.

>The first thing we discussed was why we should care about the primary intentions. Roughly the line was, "if water *is* H₂O then why do we care what water could have been. It was suggested that the reason is that if, via the PI we could determine that it was logical impossible for water to just be H₂O, then we know that we have more to look for. This, of course, is the basic line for consciousness.

Well, many of the reasons for caring about PIs are related to Frege's reasons for caring about senses. In particular, PIs seem to capture the epistemic relations between thoughts and the cognitive contents of thoughts, and are deeply related to notions such as epistemic possibility and apriority. This leads arguably to a central role for

them in philosophy of language and mind (in giving an account of content that is useful for epistemic and cognitive purposes) and in epistemology.

Within metaphysics, PIs are especially useful for analyzing modal issues. In particular, it seems that PIs are central in assessing conceivability claims, and the sense in which conceivability is a guide to possibility: if P is conceivable, it's generally plausible that there is a possible world satisfying the primary intension of C. And they are central to claims about a posteriori necessity. What's distinctive of a posteriori necessities seems to be that they have a necessary SI but a contingent PI. Given that for many issues in metaphysics we are concerned with questions about the relationship between conceivability and possibility, and about apriority and necessity, PIs seem pretty vital.

>The bulk of the rest of the discussion was on proper names again.
>Thony argues that if the PI of a name is what picks out its referent,
>then two names which pick out the same object in our world could pick
>out different objects in another world. But it seems that as a matter
>of logic, if 'a=b', where 'a' and 'b' are names, then necessarily 'a=b'
>
>I suggested that in this framework proper names are not being used as
>labels for objects. Rather they are functions which pick out objects
>of a certain description.
>
>I think Thony and I now agree that the question is whether names ought
>to work this way.

I'm a little concerned about this bit. Remember, one of the central points of the 2-D framework is that when S is necessary (in the usual sense), S has a necessary secondary intension. So what follows from the necessity of "a=b" is that "a=b" has a necessary SI, i.e. that 'a' and 'b' have the same SI. Nothing about PIs follows. In particular the claim that 'a' and 'b' have different PIs is completely compatible with the claim that "a=b" is necessary. (Just look at "Hesperus = Phosphorus", etc.)

Once can think of the PI as being what picks out reference in the actual world (depending on how it turns out) and the SI as picking out reference in counterfactual worlds (given that the actual world is fixed). I.e., PI gives reference across worlds considered as epistemic possibilities, SI across worlds considered as subjunctive possibilities. Given this, then in a case such as the above (with different PI, same SI), 'a' and 'b' will pick out the same referent in all counterfactual worlds (all subjunctive possibilities), as Kripke et al lead us to expect. All that follows from the difference in PIs it is epistemically possible that reference comes apart, i.e. there are epistemic possibilities such that if they turned out to be actual, reference would come apart, which seems to be correct in these cases.

Proper names are still being used as labels for objects in this framework, and they are still rigid designators. They pick out the same object across counterfactual worlds (considered as counterfactual). There's a sense in which actual-world reference is mediated by something a bit like a description (see previous note), but that's a far cry from saying that names here function like descriptions (which are nonrigid, etc, and have very different SIs). Witness the difference between "Hesperus" (the name) and "evening star" (the description).

I had hoped that all this was clear by now, but the above suggests that maybe there is still a little confusion. Or maybe I am just misreading the above, in which case feel free to elaborate on the issue.

>> Brad tried to provide motivation for thinking that names do in fact have
>> descriptive content. One such motivation comes from considering cases in
>> which co-referring names can't be substituted *salva veritate* (such as in
>> belief contexts). He also mentioned substitution failures in "simple
>> sentences" (see Saul 1997), such as "Lois Lane kissed Superman before she
>> kissed Clark Kent".

This is interesting indirect evidence, but I don't think the case for PIs relies on this sort of evidence. The mere epistemic possibility that Superman is not Clark Kent, and the way the reference of a term varies depending on which epistemic possibility turns out to be actual, is enough to make the case. Again, these things don't imply descriptive reference across counterfactual worlds, but they do arguably imply some quasi-descriptive content across epistemic possibilities.

Incidentally it's probably a good idea not to get too fixated on issues specific to proper names, as many of those issues (especially those of reference via linguistic causal chains) aren't all that crucial to what comes later. Natural kind cases are probably more crucial.

>We then dicussed my BIV cases. From an intuition point of view we
>were evenly split about whether or not there was really any water
>after you were pulled out of the vat. Brad suggested that it may just
>be very difficult to think of this case as anything other than a
>person jumping between two possible worlds. This, of course, is
>impossible. But it does look kind of like that.

Well, it may be more palatable to look at this as a person jumping between two very different parts of one world. A bit like going from one planet to another. In this case, going from a "virtual environment" to a real environment. Different environments in one sense, but still all part of the same world (it's a big world!).

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 02:05:16 1999
Date: Fri, 19 Feb 1999 02:05:05 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: possibility?
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Anthony writes:

>I think this is probably getting somewhat off the topic, but I have a
>general worry about making sense of secondary intensions.
>
>The 2-intension of water is H₂O, given that, in this world, we take it
>that this chemical structure explains the various properties we associate
>with water. More generally, it seems that secondary intensions are
>different from the primary intension only when there is some sort of
>reductive explanation of a particular term or thing (p62).

Well, I'd say this is often the case (especially for natural kind

terms), but it needn't be the case. E.g. "one meter" has different PI and SI, but no reductive explanation in the vicinity. Similar for "Hesperus". It's true that for natural kind terms, though, we very often identify it with some explanatory microstructure across worlds. So the SI of "water" goes with H₂O, of "tiger" goes with some bio-plus-DNA kind, of "heat" goes with molecular motion, etc. I don't think an actual explanation is required -- arguably the SI of "water" picked out H₂O in all worlds even before we had an explanation. But it's arguable that the SI of most such terms picks out the explanatory microstructure of the stuff, whatever that happens to be in the actual world, across all worlds.

>It is not, of course, directly observable that water is H₂O. Rather, it is
>the result of theory constructed from a variety of experimental results.
>Consider a possible world in which water is H₂O, but in which the natural
>laws are such that the elements behave radically different than they do in
>our world. Suppose, for example, that Hydrogen atoms form extremely
>unstable compounds when they are in the presence of other Hydrogen atoms.
>In this possible world, the dominant 'watery stuff' is XYZ. Thus, XYZ is
>the stuff picked out by the primary intension of water. XYZ behaves
>experimentally much like H₂O behaves in our world. But XYZ is not composed
>of two Hydrogen atoms and an Oxygen atom in this possible world.

OK, it seems plausible that the PI of "water" picks out XYZ in this world. The Kripkean may well say that the SI of "water" picks out H₂O, irrespective of its instability.

>It seems to me that the secondary intension of water cannot just be H₂O.
>It must also include a host of other information regarding the underlying
>physical theory that makes this identification meaningful. If there is a
>possible world in which XYZ behaves for the most part like H₂O, only X, Y
>and Z are not 2 H's and an O, it seems strange that there is any value in
>the secondary intension of water picking out H₂O, stuff that is, in this
>possible world, unstable.

Interesting point. The Kripkean often talks as if any H₂O will be water. They will then describe this world as a world in which water molecules are unstable. It does arguably make sense to consider subjunctive possibilities such as "water molecules might have been unstable", or "if water molecules had been unstable, life on earth would not have evolved". If so, then such worlds (considered subjunctively) are worlds with water molecules in them, and thus (presumably) worlds with water in them.

Still, I can see your intuition that this might not really deserve to count as "water". One could similarly consider worlds where H₂O molecules make multicolored paints that people hang on walls, or are used in brick buildings, etc. Are these worlds where paintings are made of water, etc? (Even in the subjunctive SI sense?) I can at least see the counter intuition, though I go both ways.

How to accommodate that intuition? Maybe (1) Insist that the SI of water still picks out H₂O in all worlds, but say that that the molecules only count as H₂O if they are quite a bit like our H₂O molecules with current chemistry (i.e., being H₂O requires more than just having 2 H's and an O).. Or (2) identify the SI of water with something like "H₂O, as long as it has at least some of the other chemical or surface properties it has in the actual world". Or (3) go all the way to SI = "watery stuff", which would be a more radical move, letting in XYZ as water in some counterfactual worlds; unorthodox, but some (e.g. David Lewis) have been tempted to argue that there are worlds where water is XYZ.

>If this isn't just confused, it may be that this is a reason why you say
>primary intensions are particularly important to your account.

Well, I do agree that there is an awful lot of slack in what counts as the SI of a given term. It depends some pretty debatable subjunctive intuitions about what it takes be water in a counterfactual worlds. There is also some slack in a PI, but arguably not so much -- we're somewhat clearer about what we'd count as water in different ways the actual world might turn out. This isn't the main reason I think PIs are more important than SIs (the central reason is that I think PIs tie much more closer to the epistemic notions that are at the heart of content, thought, and modality), but it may be one way of getting at the fact that SIs are most relevant in subjunctive and counterfactual contexts, about which our intuitive descriptions are pretty debatable.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 05:25:06 1999
Date: Fri, 19 Feb 1999 05:23:40 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Apriority, necessity, conceivability, possibility
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Recent discussion has been interesting, but we've drifted a little from the central topics for which the 2-D framework will be important for us. In particular, the framework is most useful for analyzing the relationship between apriority and necessity, the notion of the necessary a posteriori, the relationship between conceivability and possibility, and the issue of a priori access to modality. Of course all of these have are central to important issues in metaphysics, not least to analyzing epistemic and modal arguments in the philosophy of mind. So it's worth going over a few of these things.

Most of what's contained in this note is very important to the courset, so read it carefully.

DEFINITION OF PRIMARY AND SECONDARY INTENSIONS

I'll take a moment to officially define PIs and SIs for statements. So far we've mostly been talking about PIs and SIs for terms, but often it is whole statements that are most crucial. In effect, the PI of a statement is what you get from "composing" the PI and SI of its component terms. Note that I use "primary proposition" and "secondary proposition" for this idea in the book, but now I prefer to use "primary intension" and "secondary intension" for both.

The primary intension of a statement S is a mapping from centered worlds to truth-values. The secondary intension of S is a mapping from uncentered worlds to truth-values. They're defined as follows:

The primary intension of S maps a centered world W to the truth-value of S in W when W is considered as actual (i.e. considered as an epistemic possibility).

The secondary intension of S maps a centered world W to the truth-value of S in W when W is considered as counterfactual (i.e. considered as a subjunctive possibility, with the actual world fixed).

We can give a couple of heuristics for evaluating the PI or SI of a statement S at a world W.

PI Heuristic: If W turns out to be actual, will S turn out to be true?

SI Heuristic: Given that the actual world is the way it is, is S true in world W?

Take W = XYZ-world, S = "water is XYZ". If the XYZ-world turns out to be actual, "water is XYZ" will turn out to be true, so the PI of S is true in W. But given that the actual world is the way it is (i.e. the H2O-world), "water is XYZ" is false in the XYZ-world, so the PI of S is false in W. (I abstract away from centering issues here.)

(My favorite heuristics are actually a bit more streamlined than this, though maybe a bit less self-explanatory:

PI Heuristic: If W is the case, is S the case?

SI Heuristic: If W were the case, would S be the case?

This ties PIs to indicative conditionals (which basically force you to evaluate the antecedent as an epistemic possibility) and ties SIs to subjunctive conditionals (which make you treat the antecedent as some sort of "fantasy"). But I'll say more on this later.)

APRIORITY AND NECESSITY

The 2-D framework can be understood as making the following claims concerning the necessity and apriority of statements S.

NECESSITY THESIS: S is necessary iff S has a necessary secondary intension.

Here, I'm taking "necessary" in the standard Kripkean sense (i.e. "subjunctive necessity"; I'll argue later that this usage is not wholly uncontestable, but set that aside for now). S is necessary in the Kripkean sense precisely if it is true in all worlds considered counterfactually, which is just to say that its secondary intension is true in all worlds. I hope this claim is fairly straightforward, given the definition of a secondary intension.

APRIORITY THESIS (weak form): If S is a priori, S has a necessary primary intension.

The case for this is pretty straightforward. If S is a priori, then plausibly it will be true no matter how the actual world turns out, i.e. it will be true in any world considered as actual (considered as an epistemic possibility). That is, for an arbitrary world W, if W turns out to be actual, S will turn out to be true. That is, the primary intension of S is true in all centered worlds.

APRIORITY THESIS (strong form): S is a priori iff S has a necessary primary intension.

This makes the further claim that if S has a necessary PI, it is a priori. This isn't entirely straightforward to "prove". Necessary PI means that S will be true no matter how the actual world turns out, but it is still a leap from there to the epistemic claim that we can know S to be true. Maybe S could turn out to be unknowable, for

example. E.g., maybe some really complex mathematical truth could have a necessary PI but not be knowable a priori?

We'll come back to this issue of "scrutability" later. For now, I'd consider the weak form better established than the strong form, but I'd also say its plausible that the great majority of statements with a necessary PI seem to be knowable a priori, so that exceptions to the strong form are rare if they exist.

I'm very interested to see any potential counterexamples to any of these principles (or arguments). It's not impossible that I've overlooked something here, so feel free to jig them up.

A POSTERIORI NECESSITY THESIS: S is an a posteriori necessity iff S has a necessary secondary intension and a contingent primary intension.

This follows directly from the necessity thesis and from the apriority thesis (strong form). I'd say that this is the central claim of the 2-D analysis of a posteriori necessity. It certainly seems to fit all the Kripke examples: "Hesperus is Phosphorus", "water is H₂O", etc.

Somebody might deny this thesis (presumably by denying the strong form of the apriority thesis), asserting that there is some other sort of a posteriori necessity. I think it's at least highly plausible that such a posteriori necessities would have to have their roots in considerations quite distinct from the ones Kripke raises, though.

One reason all this is useful is that it lets us jump from epistemic claims about what is and isn't a priori to modal claims about the existence of various possible worlds. E.g. if S isn't a priori, there's a possible world where its PI is false; and even if S is an a posteriori necessity, its PI will be false in some world. If this is so, it's very important to the epistemology of modality, and to the role of epistemic arguments in metaphysics.

What makes all this possible is a central property of the 2-D framework: that it has two dimensions of semantic evaluation (two intensions), but just one space of worlds. Centering aside, PIs and SIs are evaluated at exactly the same worlds. They just give different results because of the different ways the worlds are considered (as actual or as counterfactual). Thus the worlds considered as "epistemic possibilities" for PI purposes are perfectly respectable metaphysically possible worlds: witness the XYZ-world, for example. Just one world, but two different ways of looking at it. That means that when we have an a posteriori necessity, we have a genuine metaphysical possibility in which the PI of the statement is false. That metaphysical possibility may well have consequences for metaphysics.

Again, any potential counterexamples and counterarguments are welcome.

CONCEIVABILITY AND POSSIBILITY

We can also put all this in terms of the relationship between conceivability and possibility. Conceivability is an epistemic notion. There are various notions in the area, as we'll see, but perhaps the simplest is the following. (N.B. I now prefer this way of doing things to the way I do things in the book.):

S is conceivable iff S is not ruled out a priori (i.e., if it is not a priori that not-S).

If the 2-D framework is right, then one will be able to make an inference from conceivability to possibility. If S is conceivable, not-S is not a priori, so not-S has a contingent PI (by the strong apriority thesis), so there is a metaphysically possible world in which the PI of S is true. That's not to say that S itself is "metaphysically" possible in the Kripkean sense (that sense requires a world where the SI is true), but we still have a metaphysically possible world in the vicinity. Call this the

CONPOSS PRINCIPLE: If S is conceivable, there is a possible world in which the primary intension of S is true.

We'll be looking at different versions of this claim later, and examining just what it takes to support the relevant versions. But note for now how it fits the Kripkean examples. People often say that the Kripke examples show that there can be no inference from conceivability to possibility. E.g., in the relevant sense of conceivability, "water is XYZ" is conceivable, as is "Hesperus is not Phosphorus", but neither of these is possible. But on the 2-D framework, there is at least a possibility in the vicinity: a possible world in which the PI of "water is XYZ" is true, and one in which the PI of "Hesperus is not Phosphorus" is true.

Most importantly, when you conceive (in the relevant sense) of a scenario in which water is XYZ, or in which Hesperus isn't Phosphorus, you **do** conceive of a perfectly respectable world: the XYZ-world in the first case, and a world in which the evening object isn't the morning object in the second. These worlds don't satisfy the statements in question when they're considered as counterfactual, but they do when they're considered as actual. That's no surprise: after all, one is conceiving the scenarios in question as epistemic possibilities (i.e. considering them as actual), so one would expect that the worlds that correspond to this conceivability are worlds that satisfy S when they are considered as actual. So it's a little more indirect than a simple inference from S is conceivable to S is possible, but we still have conceivability as a guide to metaphysical possibility in some sense.

MODAL RATIONALISM

Note that the two-dimensional framework fits nicely with modal rationalism: the claim that we have a priori access to modality and to the space of possible worlds. Conceivability as described above is a paradigm a priori method. And when we conceive of a scenario in the relevant sense, it appears that we really are conceiving of a possible world (the XYZ-world, or the two-star world). Of course the sense in which it's an S-world is that it satisfies S's primary intension, not its secondary intension. So we have to be careful about just how we describe the worlds in question. But still, conceivability acts as a direct guide to the space of possible worlds. So the way is opened for some sort of modal rationalism.

Of course all this depends on the central claims of the 2-D framework being correct. I think there are probably a few points at which an opponent can argue (e.g. the strong apriority thesis), but it's tricky, and again they'd have to go well beyond Kripke.

PHILOSOPHY OF MIND

All this is very relevant to epistemic arguments against materialism. One response to these arguments is to deny that there is an epistemic gap between physical and phenomenal, and assert that there is an a priori entailment. I sometimes call this "type-A materialism". It's a hard row to hoe, though, because it only seems to work if one accepts a functional analysis of phenomenal concepts, and such an analysis seems very implausible.

So perhaps the most popular response is "type-B materialism": accept that there is an epistemic gap, but deny that there is an ontological gap. This doesn't have the intuitive implausibility of the type-A view, but still holds onto materialism. Typically, one will accept that there is no a priori entailment from physical to phenomenal, but will assert that there is a necessary entailment. So the entailment from physical to phenomenal is an a posteriori necessity.

This view also usually goes along with the claim that conceivability doesn't imply possibility. So zombies (for example) are said to be conceivable but not possible, so materialism is not endangered.

If the central claims of the 2-D framework are correct, this isn't so simple. The type-B materialist holds that "P $\rightarrow$ Q" is an a posteriori necessity, where P is the physical truth about the world and Q is a phenomenal truth. On the 2-D framework, it follows that "P $\rightarrow$ Q" has a contingent PI, so there is a possible world where the PI of P is true, but the PI of Q is false. Under a bit more analysis, the existence of this possible world is highly threatening to materialism.

Similarly, if zombies are conceivable, it follows that there's a possible world in which at least the PI of "there are zombies" is true. From here (under a bit of analysis) one can plausibly get to a world physically identical to ours but different overall, so that materialism comes out false.

The basic reason for all this, of course, is that deep down the Kripkean framework hasn't really broken the link between conceivability and possibility, it has just recast it a little. But we've seen that even Kripke's own results rely on the method of conceivability: conceive of a world, then make sure you describe it correctly. So it arguably supports rather than opposes a modal rationalism at the deep level. If so then the Kripkean framework gives much less support to the type-B materialist than is often supposed.

We'll be going into the full details of how this applies to materialism etc next week, so we can save a detailed analysis of the consciousness issues for then. But it's important to see how the 2-D framework might have significant metaphysical consequences, and how those rest pretty straightforwardly on the central theses of the framework.

For now, I'd be interested to see some discussion of those theses, and of the relationship between apriority and necessity vis-a-vis PIs and SIs, etc. All comments are welcome.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 18 08:10:40 1999
Date: Thu, 18 Feb 1999 09:09:39 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>

Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

>

A quick comment on Dave's comment on my comment on Josh:

[For Dave's comments, and some of my original comments, see below.]

The point I was making about glumpy stuff was this. In Josh's scenario, I'm told (and believe) that I've just left a Vat. Presumably I will also believe that none of the stuff behind my past experiences is the same sort of stuff as the stuff behind my current experiences. In this situation, I want to suggest that we shouldn't reidentify my in-Vat <"water"> concept with my post-Vat water concept, *even if we're picking out concepts in terms of their primary intensions.* It is true that both in the vat and out of the vat, I have a concept that is closely connected with my phenomenal concept <watery stuff>, and it's a natural kind concept too. Call these two concepts 'Water1' and 'water2' (These are just two names, I'm not yet supposing that these are two distinct concepts). When, out of the Vat, I use water2, I don't think of it as applying to the same sort of stuff that water1 applied to. And that, it seems to me, is enough to say that its primary intension is different.

Here's an analogy. Suppose I have two concepts, <glumph> and <water>. I think that glumph and water are distinct natural kinds. Both act in the same (watery) way, indeed, I can't tell them apart. Almost nobody can tell them apart. (Perhaps nobody now living can tell them apart. It's just part of our mythology that there were people - the shabby pedagogues - who could tell them apart.) But I know that the stuff around here is glumph. Still, as far as I'm concerned, I interact with glumph pretty much the same way I would with water. Now, suppose that I move to an area where the watery stuff is not glumph but water. *If I know that this is the case*, I will use a different concept, my <water> concept rather than my <glumph concept>, to refer to the watery stuff around me.

I want to say that this case is precisely analogous to Josh's BIV case in which he believes that he was a BIV. It may look like he is still using his water concept, but he's not. He's using a concept that has the same surface structure, but there's more to PI than surface structure. At least for Natural Kinds, one has to think that one is talking about the same sort of stuff in order for "two" primary intensions to be the same.

Well, maybe.

> There's also the special case where I'm told I've shifted, and where I
> make a new stipulation about my concepts so my beliefs don't come out
> useless. That would plausibly involve a "different concept" in a
> reasonably strong sense, but that also seems to be something of a
> special case.
>
> Tim raises some worries about concept identity:
>
> >But should we think that there is any sense in which VAT-Josh
> >and post-Vat Josh have a concept, i.e. <water> that can be tracked from
> >the VAT to after the VAT? I wonder. Both in the Vat and after the vat Josh
> >has a concept that is related to <watery>, but that doesn't seem
> >sufficient to justify the claim that he has a single concept that survives
> >the experience from VAT to actual world, that we can reidentify as the
> >same concept. After all, both my concept of a certain chemical substance

> >(say, glump - suppose that glup is watery) and my concept <water> are
> >connected to <watery stuff>, but that doesn't give me any incentive to
> >identify <glump> with <water>.
>
> I presume you're talking about the vatted concept and the
> long-post-vat concept. In favor of "same concept" is that they have
> the same PI, and a continuous history of use between them. Against
> "same concept" is that they have a different referent and SI. What
> answer one gives will partly depend on whether one individuates
> concepts by PI, SI, referent, or history of use. I think this is
> somewhat terminological. I have some temptation toward "same
> concept", given the centrality of PIs, but then the difference in
> referent and SI is pretty significant, too. As for water/glump, I'm
> not sure how to imagine the case -- do they pick out the same
> referent, or not? If the same referent, presumably we have same SI,
> different PI, so the reverse of the above. If a different referent,
> presumably we have different PI and SI, which is worse (presumably
> even if the PI of "glump" involves "watery", it will be in a different
> way, as it isn't a priori that water is glump). Either way, one
> relevant difference is that the vat case involves sameness of PI and
> continuity of use, whereas this case doesn't.
>
> >To sum: supposing the individualistic spin on water, in which the primary
> >intension is <whatever has been causing *my* watery experiences> we (and
> >Josh) might still be tempted to identify his in-Vat watery concept with
> >his post-Vat watery concept, even though it refers to a different stuff,
> >because many of the inferential connections that he learnt while in the
> >vat still hold true in the actual world. If post-Vat Josh thinks of his
> >envatted experiences as including water thoughts, then it is clear that he
> >must have a water concept that we can trace from the Vat to the real
> >world.
>
> I think I agree with this, more or less. There are actually two
> issues: (1) sameness of concept pre- and post-vat, and (2) whether the
> concept picks out the actual stuff. That the "inferential connections
> still hold true in the actual world" seems to be getting at (2) rather
> than (1). In my terms, I'd put it by saying that as long as the
> actual world isn't *too* different from the vat world, actual watery
> stuff can come to satisfy the PI of my "water" concept (though
> possibly with a time lag). I'm inclined to think that (1) doesn't
> turn so much on the actual facts, but more on facts about the
> subject's psychological structure (modulo issues about individuating
> concepts by referent). That seems to fit with your last sentence
> above.
>
>
> (3) Tim raises the case where I'm shifted to Twin Earth in my sleep.
> As I said in the last message, I think Tim is right that my "water"
> won't initially refer to XYZ, though it may after a couple of years.
> I think the moral of that is that the PI of "water" anchors one to
> one's past environment, as Tim suggests. Basically because it picks
> out (at least in part) the stuff that one has causally interacted with
> in the past. In fact I think that for many of these cases, the
> present environment is mostly irrelevant (except to determining
> whether one's beliefs are now true or false, etc). The future
> environment even more so, though maybe one could come up with cases
> where it matters.
>
> All thoughts and reactions are welcome!
>
> --Dave.

>

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 18:10:24 1999
Date: Wed, 17 Feb 1999 18:07:56 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Here are a few thoughts on Josh's "water" case. In this message I'll give my own analysis, and in the next I'll give some thoughts on others' analyses.

The first thing to get straight on is whether one thinks the judgments in question are true or false, before and after the vat.

Case 1: I think "water is H2O" in the vat. I escape from vat, find a similar-seeming world, except that the watery stuff is XYZ.

Case 2: I think "water is H2O" in vat, until a within-vat revolution convinces me of "water is XYZ". I escape from vat, find a similar-seeming world, and the watery stuff there is XYZ.

Q1: What (if anything) do I refer to as "water" pre-escape?
Q2: What (if anything) do I refer to as "water" post-escape?
Q3: Is my statement "water exists" true or false pre-escape?
Q4: Is my statement "water exists" true or false post-escape?

There are various things to say here. It's probably simplest to first set aside the belief changes re XYZ and H2O, and think about what "water" might refer to pre- and post-vat in an ordinary BIV case. Here some distinct issues peculiar to BIV cases come up. For example, many would say that while in the vat, my term "water" doesn't really refer to anything, and that my thought "water exists" is false. After all, isn't a brain in a vat a classic case of someone who is deceived about the external world?

One could accommodate this line in the 2-D framework by saying e.g.: (1) the PI of "water" picks out something like "the watery stuff in the environment that the being at the center has been causally related to" (not an implausible analysis of my "water" PI), and (2) if I am a BIV, my centered world (i.e. my actual world, not merely the world-as-I-believe-it-to-be) doesn't have anything that satisfies that PI. (There's no watery stuff causally related to me in my actual environment.) So the PI will pick out nothing, and my vatted thought "water exists" will be false.

On this line, what will happen when I come out of the vat? Well, arguably when I emerge for the first time into the "real world", my term "water" doesn't yet refer to actual water, because there's no causal connection yet. But after interacting with the environment for

a while, the causal connection will get going, and perhaps eventually my term "water" will come to refer to the actual stuff (H₂O). Note that all that will be compatible with the PI analysis above: nothing will satisfy the PI at stage 1 (centered on me in vat) at stage 2 (centered on me just after escape), but something may satisfy it at stage 3 (centered on me well after escape), as by then the center will have plenty of causal relations with some watery stuff.

Another line that some philosophers take is that while in the vat, my terms like "water" refers to some chemicals or nutrients in the vat around the brain, because these are what I'm causally connected to. I don't find this very plausible myself, but if one takes this line, one could accommodate it by dropping "watery stuff" from the PI above, and by saying that the PI picks out roughly "the stuff in the environment that is causally responsible for the being at the center's use of the word 'water'". In this case, we'll have "water exists" true at stage 1 (it picks out chemicals), arguably false at stage 2 (it still picks out chemicals, but I suppose that could be true since the chemicals still exist), and true again at stage 3 (by now, the PI picks out real water).

Alternatively, one could be tempted by the idea that while in the vat, one's term "water" refers to something in one's "virtual world", or something like that, so that one's vatted thoughts "water exists" is true. This has the problem of making it hard to express how the BIV is deceived, and could even end up leading to some sort of idealism or phenomenalism. But there is arguably something to the intuition that there's some sense in which the BIV's "water is H₂O" is truer than "water is XYZ" (assuming the virtual world is an H₂O-world). Maybe one could accommodate this by saying there's two ways to read such thoughts, a strong way in which they all come out false, and a "weak" way where some come out true, or some such. E.g., the "weak" way might involve prefixing "According to the virtual world", or might have a PI involving "seeming-objects which seem to be causally connected to me", or some such. This would be a tricky row to hoe, but maybe there are possibilities. In this case, one could say that reference of "water" gradually shifts after release, just as above.

Anyway, I think one can factor away from these tricky BIV issues for considering Josh's main point. I'm not sure whether Josh's point would work equally well if the vat world were replaced by another planet in which one is brought up, and that brought to earth. That would probably give similar results at least to line 3 above, and arguably to line 2, though maybe not to line 1. But the crucial claims about how "water" refers after release may be similar each way.

What to say about case 1, where I believed "water is H₂O", and find out post-vat that the watery stuff is XYZ? I'd be tempted to say, as in the cases above, that immediately post-release my term "water" still refers to whatever it referred to just pre-release: either nothing, or some chemicals, or perhaps some virtual H₂O in the virtual world. So my statement "water exists" then will be false, or in any case if true won't be made true by the actual watery stuff. But after a while of interacting with the real environment, my term "water" might well come to refer to the XYZ, and my claim "water exists" will be made true by the actual environment.

I think that's so despite the fact that I believed in the vat that water was H₂O. After all, within the vat that belief was potentially malleable (if people had told me the watery stuff was XYZ, I'd have accepted that water was XYZ), so the same ought to be the case outside

the vat. In fact, I think that even if no-one outside the vat ever tells me that water is XYZ, my term "water" may eventually come to refer to the XYZ by virtue of my causally interacting with it, etc. In this case, I'll eventually have a false belief that "water is H₂O" but a true belief that "water exists".

If that's how things so, I think it's compatible with the PI staying constant throughout. E.g. if the PI is "the watery stuff in the environment that the being at the center has been causally related to", this seems to give the right results: it picks out nothing at stage 1, nothing at stage 2, XYZ at stage 3. And that's so irrespective of my belief changes.

I'd be tempted to say something similar about case 2. Here, I think there still needs to be a gradual reference-shift from vat-reference to actual reference. Even if one thought "water is XYZ" in the vat, and actual watery stuff is XYZ, it's not clear that one's thought (immediately post-escape) "water is XYZ" is true, or at least it's not clear that it's made true by the actual XYZ. It takes some interaction for the term "water" to pick out the actual stuff.

[A side note: It may be that if the scientists tell you immediately "you were in a vat, but now you're out, and fortunately the real world is a lot like you thought the vat world was", you might decide to make a new stipulation that all your words will refer to actual things rather than vat things, rather than keeping a lot of terms around that don't refer to anything useful. That's tricky in some cases: e.g. re "my wife", or "Ethel", say, does one really want to say that the term refers to a person in the actual world who you've never strictly speaking met before? But I suppose one could try to stipulate that at least some terms refer to their qualitative counterparts. If that's so, then "water" might immediately pick out the respective watery stuff.

As for what to say if there are a few small differences between vat world and actual world (e.g. they replaced H₂O by XYZ), that's awfully tricky. I guess it depends on one's stipulation. One might stipulate "in case of any differences, my term doesn't refer", or one might stipulate "in case of any differences, my term refers to whatever is superficially similar enough", or some such. In the first case, my post-escape "water exists" might at first be false; in the second; it might be true. But that all depends on just what stipulation one makes. Such a stipulation might well change one's term's PI a little bit, as any new stipulation will.

Probably it's easiest to think about these things in a case where the scientists don't tell you you've escaped, so one isn't faced with the immediate wrenching discovery that forces you to reevaluate your whole conceptual scheme. Instead, maybe do things so that you don't know you've undergone such a radical shift. Even so, we can evaluate truth and reference, and we can still consider what happens if there turn out to be small changes in the environment. That way the "stipulation" aspect is removed.]

Anyway, that's how I'd be tempted to analyze these cases. I think it's plausible on most ways of doing things that one's PI stays constant, but its referent may gradually shift, in effect due to a change in the centered world in question (or at least, a change in the location of the center). The one case where my PI may change is the case where I know that I've shifted environments, and decide to restipulate the use of my terms; but that's a very special sort of change.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 17:24:53 1999
Date: Wed, 17 Feb 1999 18:24:00 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on Josh's BIV case and on de se representation. . .

The BIV case is a nice case, and a difficult one. I had an answer, but talked myself out of it. Here's another stab. This is in three parts, all rather convoluted.

(1) Part of what is up for issue is whether the primary intension of our concepts like <water> are egocentrically indexical, or whether they are more communal. Suppose that Josh is the only BIV, and that he thinks of water as <the stuff that causes (most of) most of *my* watery experiences>. On this construal of the 1-intension, Josh's water concept refers to certain types of electrodes, or computer programs, or whatever it is that causes most of his watery experiences. (Why do we describe it as his "water concept"? Cos it's connected to his watery experiences, which are connected to our water concept.) This, I take it, is the Putnamian position on BIVs. On a *communal* interpretation of the 1-intension of water, it is <the stuff that causes most of the watery experiences of most of the folks around here.> (There are lots of indexicals in it). Now, whether this concept still refers to the electrodes or programs of whatever depends on whether Josh is the only BIV around. If the vast majority of the folks around here are also BIVs, then perhaps one could argue that even now it still refers to these electrodes. (Of course, the problem is that Josh isn't communicating with any of these - presumably. So maybe the communal condition can't get a toe-hold in giving the referent of an individual BIV's concept.)

But let's suppose that the 1-intension of Josh's <water> is communal, and that he's the only BIV around here. Then, his thoughts about water would be mostly mistaken cos they are about water, and he isn't connected to water in the right way. Although of course his belief that water is H₂O would be true if the stuff that causes most people's watery experiences is H₂O. Of course, it may not always be easy to tell whether one of his thoughts is true or false, because it may not be clear whether it is about water. Suppose that he thinks that he is having a shower: 'Ah, a lovely shower, this stuff sure is hot' and such like. This might be a true thought if the stuff that is causing his experiences is indeed hot. But insofar as he thinks that the stuff that is hot is the same kind of stuff that everyone else (i.e. non-BIVs) thinks is hot when they are taking a shower, his thought is false. Similar comments applying to his thought, 'It sure is good to run naked through the water of this Bondi surf.' This is true, insofar as it is good to run naked through the Bondi surf, but of course it is false in that Josh-the-BIV is not running anywhere, least of all running naked through the Bondi surf.

(2) Let's push the Putnamian line for a minute and see where it goes. Take Josh-the-BIV's water concept to refer to whatever causes his watery experiences. Post-vat Josh's water concept also refers to the stuff that causes his watery experiences. These two concepts (mental representations) are similar in that they are both related to watery experiences, that is, they are both object concepts that are related to the *phenomenal concept*

<watery>. But should we think that there is any sense in which VAT-Josh and post-Vat Josh have a concept, i.e. <water> that can be tracked from the VAT to after the VAT? I wonder. Both in the Vat and after the vat Josh has a concept that is related to <watery>, but that doesn't seem sufficient to justify the claim that he has a single concept that survives the experience from VAT to actual world, that we can reidentify as the same concept. After all, both my concept of a certain chemical substance (say, glump - suppose that glup is watery) and my concept <water> are connected to <watery stuff>, but that doesn't give me any incentive to identify <glump> with <water>. In order to identify the concept that we are calling Josh's pre-Vat water concept with his post-Vat water concept it needs to be the case that most (much?) of what he has learnt about water in the VAT holds true in the actual world. Does envatted Josh's knowledge of the microstructure of water play an important role here? I don't see why. On an inferential role account of concepts, beliefs about the microstructure of a substance might be just one sort of inferential node among many. If the microstructure of XYZ is such that it behaves in almost all of the same ways that H₂O behaves, then we might be inclined to say that Josh has learnt that water is XYZ, or whatever it is in the actual world. (Of course, XYZ can't behave in all of the same ways that H₂O does, otherwise we would be able to discover that some stuff is H₂O rather than XYZ.)

To sum: supposing the individualistic spin on water, in which the primary intension is <whatever has been causing *my* watery experiences> we (and Josh) might still be tempted to identify his in-Vat watery concept with his post-Vat watery concept, even though it refers to a different stuff, because many of the inferential connections that he learnt while in the vat still hold true in the actual world. If post-Vat Josh thinks of his envatted experiences as including water thoughts, then it is clear that he must have a water concept that we can trace from the Vat to the real world.

(3) Here's a similar case to Josh's that puts it in a communal setting. Suppose that we are all shifted to Twin Earth in our sleep. Twin Earth is as the travel brochures describe it: it's the same as earth, except the predominant watery stuff in the environment is XYZ. After a couple of days we would discover that "this stuff is XYZ". I guess most of us would say, "Huh, I guess that water isn't H₂O after all, it's XYZ." But it seems that we would all be mistaken. Water is H₂O, but this ain't water. Of course, after a couple of years on twin earth, 'water' would come to refer to XYZ. So it seems.

Well, I'm not sure what the point of this scenario is. Maybe it raises the question: 'Is the 1-intension of <water> "the stuff around here that causes most of our watery experiences", or is the 1-intension of water "the stuff that is causing and has caused most of the watery experiences around here at the moment and in the past." The suggestion is that the 1-intension of natural kind terms anchors them to the past (and maybe the future).

Tim

University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 20:31:28 1999
Date: Wed, 17 Feb 1999 20:29:40 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on analyses of Josh's case by Josh, Erik, Erik, Angela, and Tim. N.B. This presumes the previous message with my own analysis.

Josh wrote:

>Now here is the puzzle. Primary intentions pick out "watery stuff" in
>the actual world. My intuition is that the actual world in both cases
>is the post-vat world. However, if I were the one pulled out of the
>vat in case 1 I'd say that there is no 'water' in the actual world.
>There is just watery stuff. But if I were pulled out of the vat in
>case 2, I'd say there is water in the actual world.
>

Like others, I'm not sure I share your intuitions here. Maybe there's some reason for saying that post-escape, your term "water" doesn't refer to anything in the actual world. But I think one can argue when it does begin referring to something in the actual world, it will refer just as soon in each case. After all, the only difference between the cases is that in one I believe "water is H2O" and in the other I believe "water is XYZ". This might make a difference to reference if these were "core beliefs" for me and unrevisable, or some such. But in fact it seems that these beliefs are quite revisable, a posteriori beliefs, as reflected by the fact that upon "discovering" in the vat that watery stuff is XYZ, one would say "water is XYZ" is true, not that water doesn't exist. I think the same goes outside the vat. Insofar as my word refers to stuff outside the vat at all, it's in virtue of its role as watery stuff, and it doesn't matter whether it turns out to be XYZ or H2O.

>If you share these intuitions, then there seem to be two possibilities:
>1) My primary intention of water includes the chemical make up of
>water or 2) Primary intentions do not fix reference in the actual
>world but fix reference in some other way.

Hmm. I don't share the intuitions, but perhaps one can make sense of your intuitions by saying: the PI of water picks out "whatever has the same structure as the stuff I've been related to in my past environment", and going for a view (the third line in my previous post) on which in the vat environment, you're related to virtual-H2O and virtual-XYZ in case 1 and case 2 respectively. Then post-vat, out in the real XYZ-world, your "water exists in this environment" might come out true in case 2 but not in case 1 (because XYZ has the same structure as virtual-XYZ, but not as virtual-H2O). That would seem to give the results of your intuitions, though I don't know whether (a) it gets at what underlies your intuitions, or (b) it's plausible.

[N.B. If things work this way, it's immediately post-escape that the changes will show up, rather than later. I.e. even before causal interaction with the new stuff, my "water exists around here" will be true in case 2 rather than case 1. With enough causal interaction, reference might still shift so that the claim will be true in both cases.]

This analysis may get at something about what you're saying re "PIs don't fix reference in the actual world". I'd say that even on this analysis, they do fix reference in the actual world, but the actual world is a 4-D spacetime manifold, i.e. it includes all the facts about the past. And here, your PI will fix reference partly by appealing to facts about your past environment (i.e., the way things are in the past environment of the being at the center of the world) rather than your current environment.

(It's plausible that something like that applies to many cases. E.g., if I am transported to Twin Earth without knowing, then presumably my immediate claim "there is water around here" will be false, even though there is watery stuff (in the loose sense) around there. Essentially, this is because the PI of water fixes reference to something like "the watery stuff I have causally interacted with", which involves reference to my past environment as well as my current environment. Of course if I live on Twin Earth long enough, my reference may shift enough so that "there is water around here" will be true; that's also predicted by this PI.)

Looking at your (Josh's) more recent second message, maybe something like this analysis fits.

>Regarding the responses to my BIV cases, I'd like to make a few
>clarifications. There are two ways that I see of accounting for my
>intuitions. The first is to claim that the chemical make up of water
>is part of my PI. The second is to claim that my PI is just like
>everyone else's BUT PIs don't map centered worlds considered as actual
>onto referents.

>
>I think the second of these claims is the more likely. The idea would
>be that PI's perhaps map centered worlds thought to be actual during
>concept formation onto referents. (I already see several problems
>with this definition, but that is the basic idea.) Since vat-Josh
>formed his concepts while thinking that vat-world was actual, he would
>then map water to clear drinkable liquid in vat-world. And in
>Vat-world water is H₂O.

Well, one problem with what you say above is that if I grow up (in the actual world) thinking water is XYZ, then this formula will predict that my claim "water exists" is false, which doesn't seem right. And more generally it seems that it will be hard to have false beliefs while one is growing up, which seems problematic. Maybe it's better to say that something special is going on with the vat case -- these aren't just any old false beliefs about the world, but a whole virtual environment, one that is so self-consistent and central that it deserves to count in some sense as my "actual environment" while I'm in the vat. If so, then one can argue that my vat terms fix reference to things in my virtual environment, and that even after leaving the vat, they still involve the way things were in the vat environment.

Note again that this doesn't involve ignoring the actual world, but it does involve concentrating on a particular aspect of it, i.e. my past environment, which in this case will be my past virtual environment (which of course is in some sense part of the actual world).

On this way of doing things, the escape from vat to world is just like going from Earth to Twin Earth (in case 1) or like going from Twin Earth 1 to Twin Earth 2 (in case 2). Either way, one shifts environment, and one's initial reference is fixed to the stuff in one's original environment, so the truth of one's immediate-post-move

claim "there is water around here" depends on whether there is anything around here with the same structure as the stuff in one's past environment. So true in case 2 but not in case 1. That seems to accommodate your intuitions, even with a constant PI that fixes reference via the actual world (including past environment).

Anyway, note that one this way of doing things, what makes the differences between the cases isn't just that one had different beliefs in the past about one's environment -- it's that in some sense, one actually had different environments in the two cases (albeit virtual environments, in Josh's version). To see the difference, think about what we'd say if (a) the virtual environment "really had" XYZ structure in it (i.e. if you'd looked under a virtual microscope, that's what you'd have found), but (b) you believe the whole time that there is H₂O there (you never look under a microscope). I predict that in this case, Josh will say (or ought to say) that "water is H₂O" is false when thought in the vat, and false after escape, while "water is XYZ" is true both times. That suggests that it's the past environment that matters, not the past beliefs.

Of course one can question (a) whether a "virtual environment" is really enough of an environment to make concepts refer and to make beliefs true, and even then one could question (b) whether the "real" XYZ environment really qualifies as having the "same sort of stuff" that is present in the virtual environment (one might argue that there are deep differences between virtual XYZ and real XYZ, so that the "virtual" concept won't pick out the "real" stuff after escape). But assuming one can make a case for the appropriate answer on those two questions, then Josh's position will come out as reasonable, and the PI above seems to capture roughly what's going on.

Re Erik L's, Angela's, and Erik H's messages: I think my intuitions largely agree with these, i.e. that there's no major difference between the two cases: insofar as case 2 "water" picks out the actual stuff post-vat, so does case 1 "water". (Although after going through the above, I have to say that I feel at least some of the tug of Josh's way of doing things, at least on the "virtual environment" reading.) I think one still needs to distinguish what happens immediately after release from what happens after a period of adjustment, though. I'm tempted to say that there is an element in the PI of "water" requiring a previous causal connection to the stuff, so it doesn't refer to the actual stuff immediately (in either case), but only after a while.

If one doesn't agree with the causal requirement, one could go with a looser "the watery stuff around here now" PI, so that I'll come out referring in both cases immediately after escape. But I think the causal requirement is independently plausible.

I do largely agree with what Angela says here:

>I understand the temptation, especially given the amount of scientific
>information that is part of common knowledge, to say that H₂O really is
>part of the primary intension of water, but, the very fact that I (and
>presumably others) can make sense of counterfactuals like, "Water could
>have been other than H₂O," or "Water might not have been drinkable by
>humans", or "Water might not have had the property of appearing blue in
>daylight," etc...strongly suggests that almost nothing of what we know a
>posteriori of water is contained within the primary intension.

One thing to note is that "could have" and "might have"

counterfactuals are often ambiguous between a PI and a SI reading, i.e. between (broad) epistemic possibility and Kripkean "metaphysical" or "subjunctive" possibility. I'm not 100% sure which (or both) you're meaning to appeal to here. Arguably one needs both to make best sense of these counterfactuals.

If one accepts Kripke's subjunctive intuition, "water could have been other than H₂O" will be true in the epistemic (PI) sense but not the subjunctive (SI) sense. As for "water might not have been blue", etc, this is most straightforward for me to make sense of on the subjunctive (SI) sense: one pictures the actual stuff, i.e. H₂O, having a different appearance. On the epistemic (PI) sense, this is harder to make sense of, since water's appearance properties are relatively close to its PI, i.e. are "relatively a priori" of the "water" concept. But presumably one can still make sense of them by considering certain strange (broad) epistemic possibilities, e.g. on which there has been an optical illusion, or water's surface properties are very different from how one thinks they actually are. That reflects the fact that the beliefs in question aren't so "core" as to be unrevisable. As one builds in more and more core surface properties of water, though, this gets harder to do, and the epistemic "might have" claim becomes harder to make sense of.

Re Tim's message:

(1) Tim raises a possibility I didn't mention in the last message, that my vatted "water" concept may be deferential, fixing reference via my community. Still, I'd argue that it seems wrong to say that I defer to the people out there in the real world outside the vat, and it also seems wrong to say that I defer to other BIV's who happen to be out there. (It doesn't seem right to say that those people's reference makes a difference to my reference.) If I'm deferring, it's to people in my "virtual" world, the ones I think I "got" the term from. (I suppose it could be that those virtual people in my world are "avatars" of "real" people or of other BIVs, so I'd be deferring indirectly to those people, but I'll set that possibility aside.)

If that's so, then I don't think deference makes an essential difference to the picture. If the non-deferential concept picks out nothing, or chemicals, or virtual XYZ, so will the deferential concept, on this way of looking at things. The main difference will be the precise nature of the "causal chain" connecting me to the referent, i.e. on whether it needs to go through other "virtual people", or not.

(2) Tim also raises the question of in what sense I'll have the same concept inside and outside the vat. This seems most straightforward in the case where I don't know I've shifted. In that case, there seems to be the same sort of continuity that one would find if I go from one world to another without knowing. I'll deploy a lot of concepts, and they certainly won't be a whole new repertoire that I've acquired overnight -- rather, they'll be the old concepts (although a lot of my resulting beliefs may now come out false). It may be that eventually reference will shift, so that in a certain sense I'll have different concepts -- I'll at least have concepts with a different referent and SI, though not a different PI -- but that's a slightly different phenomenon.

There's also the special case where I'm told I've shifted, and where I make a new stipulation about my concepts so my beliefs don't come out useless. That would plausibly involve a "different concept" in a

reasonably strong sense, but that also seems to be something of a special case.

Tim raises some worries about concept identity:

>But should we think that there is any sense in which VAT-Josh
>and post-Vat Josh have a concept, i.e. <water> that can be tracked from
>the VAT to after the VAT? I wonder. Both in the Vat and after the vat Josh
>has a concept that is related to <watery>, but that doesn't seem
>sufficient to justify the claim that he has a single concept that survives
>the experience from VAT to actual world, that we can reidentify as the
>same concept. After all, both my concept of a certain chemical substance
>(say, glump - suppose that glup is watery) and my concept <water> are
>connected to <watery stuff>, but that doesn't give me any incentive to
>identify <glump> with <water>.

I presume you're talking about the vatted concept and the long-post-vat concept. In favor of "same concept" is that they have the same PI, and a continuous history of use between them. Against "same concept" is that they have a different referent and SI. What answer one gives will partly depend on whether one individuates concepts by PI, SI, referent, or history of use. I think this is somewhat terminological. I have some temptation toward "same concept", given the centrality of PIs, but then the difference in referent and SI is pretty significant, too. As for water/glump, I'm not sure how to imagine the case -- do they pick out the same referent, or not? If the same referent, presumably we have same SI, different PI, so the reverse of the above. If a different referent, presumably we have different PI and SI, which is worse (presumably even if the PI of "glump" involves "watery", it will be in a different way, as it isn't a priori that water is glump). Either way, one relevant difference is that the vat case involves sameness of PI and continuity of use, whereas this case doesn't.

>To sum: supposing the individualistic spin on water, in which the primary
>intension is <whatever has been causing *my* watery experiences> we (and
>Josh) might still be tempted to identify his in-Vat watery concept with
>his post-Vat watery concept, even though it refers to a different stuff,
>because many of the inferential connections that he learnt while in the
>vat still hold true in the actual world. If post-Vat Josh thinks of his
>envatted experiences as including water thoughts, then it is clear that he
>must have a water concept that we can trace from the Vat to the real
>world.

I think I agree with this, more or less. There are actually two issues: (1) sameness of concept pre- and post-vat, and (2) whether the concept picks out the actual stuff. That the "inferential connections still hold true in the actual world" seems to be getting at (2) rather than (1). In my terms, I'd put it by saying that as long as the actual world isn't *too* different from the vat world, actual watery stuff can come to satisfy the PI of my "water" concept (though possibly with a time lag). I'm inclined to think that (1) doesn't turn so much on the actual facts, but more on facts about the subject's psychological structure (modulo issues about individuating concepts by referent). That seems to fit with your last sentence above.

(3) Tim raises the case where I'm shifted to Twin Earth in my sleep. As I said in the last message, I think Tim is right that my "water" won't initially refer to XYZ, though it may after a couple of years. I think the moral of that is that the PI of "water" anchors one to

one's past environment, as Tim suggests. Basically because it picks out (at least in part) the stuff that one has causally interacted with in the past. In fact I think that for many of these cases, the present environment is mostly irrelevant (except to determining whether one's beliefs are now true or false, etc). The future environment even more so, though maybe one could come up with cases where it matters.

All thoughts and reactions are welcome!

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 18 16:44:56 1999

Date: Thu, 18 Feb 1999 12:47:46 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Josh Cowley <josh@MATH.ARIZONA.EDU>

Subject: Minutes from Tuesday

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Here are the minutes from Tuesday. I thank Brad for filling in some details.

The first thing we discussed was why we should care about the primary intentions. Roughly the line was, "if water *is* H₂O then why do we care what water could have been. It was suggested that the reason is that if, via the PI we could determine that it was logical impossible for water to just be H₂O, then we know that we have more to look for. This, of course, is the basic line for consciousness.

The bulk of the rest of the discussion was on proper names again. Thony argues that if the PI of a name is what picks out its referent, then two names which pick out the same object in our world could pick out different objects in another world. But it seems that as a matter of logic, if 'a=b', where 'a' and 'b' are names, then necessarily 'a=b'

I suggested that in this framework proper names are not being used as labels for objects. Rather they are functions which pick out objects of a certain description.

I think Thony and I now agree that the question is whether names ought to work this way.

> Brad tried to provide motivation for thinking that names do in fact have
> descriptive content. One such motivation comes from considering cases in
> which co-referring names can't be substituted salva veritate (such as in
> belief contexts). He also mentioned substitution failures in "simple
> sentences" (see Saul 1997), such as "Lois Lane kissed Superman before she
> kissed Clark Kent".

> We also briefly discussed again the matter of what the primary intensions
> of names typically are (such as in the London/Londres case discussed
> before).

We then discussed my BIV cases. From an intuition point of view we were evenly split about whether or not there was really any water after you were pulled out of the vat. Brad suggested that it may just be very difficult to think of this case as anything other than a person jumping between two possible worlds. This, of course, is impossible. But it does look kind of like that.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 17 12:01:52 1999

Date: Wed, 17 Feb 1999 12:59:13 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: possibility?
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I think this is probably getting somewhat off the topic, but I have a general worry about making sense of secondary intensions.

The 2-intension of water is H₂O, given that, in this world, we take it that this chemical structure explains the various properties we associate with water. More generally, it seems that secondary intensions are different from the primary intension only when there is some sort of reductive explanation of a particular term or thing (p62).

It is not, of course, directly observable that water is H₂O. Rather, it is the result of theory constructed from a variety of experimental results. Consider a possible world in which water is H₂O, but in which the natural laws are such that the elements behave radically different than they do in our world. Suppose, for example, that Hydrogen atoms form extremely unstable compounds when they are in the presence of other Hydrogen atoms. In this possible world, the dominant 'watery stuff' is XYZ. Thus, XYZ is the stuff picked out by the primary intension of water. XYZ behaves experimentally much like H₂O behaves in our world. But XYZ is not composed of two Hydrogen atoms and an Oxygen atom in this possible world.

It seems to me that the secondary intension of water cannot just be H₂O. It must also include a host of other information regarding the underlying physical theory that makes this identification meaningful. If there is a possible world in which XYZ behaves for the most part like H₂O, only X, Y and Z are not 2 H's and an O, it seems strange that there is any value in the secondary intension of water picking out H₂O, stuff that is, in this possible world, unstable.

If this isn't just confused, it may be that this is a reason why you say primary intensions are particularly important to your account.

Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 01:44:18 1999
Date: Fri, 19 Feb 1999 01:44:04 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Minutes from Tuesday
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

A few comments on the minutes.

>The first thing we discussed was why we should care about the primary
>intentions. Roughly the line was, "if water *is* H₂O then why do we
>care what water could have been. It was suggested that the reason is
>that if, via the PI we could determine that it was logical impossible
>for water to just be H₂O, then we know that we have more to look for.
>This, of course, is the basic line for consciousness.

Well, many of the reasons for caring about PIs are related to Frege's reasons for caring about senses. In particular, PIs seem to capture the epistemic relations between thoughts and the cognitive contents of

thoughts, and are deeply related to notions such as epistemic possibility and apriority. This leads arguably to a central role for them in philosophy of language and mind (in giving an account of content that is useful for epistemic and cognitive purposes) and in epistemology.

Within metaphysics, PIs are especially useful for analyzing modal issues. In particular, it seems that PIs are central in assessing conceivability claims, and the sense in which conceivability is a guide to possibility: if P is conceivable, it's generally plausible that there is a possible world satisfying the primary intension of C. And they are central to claims about a posteriori necessity. What's distinctive of a posteriori necessities seems to be that they have a necessary SI but a contingent PI. Given that for many issues in metaphysics we are concerned with questions about the relationship between conceivability and possibility, and about apriority and necessity, PIs seem pretty vital.

>The bulk of the rest of the discussion was on proper names again.
>Thony argues that if the PI of a name is what picks out its referent,
>then two names which pick out the same object in our world could pick
>out different objects in another world. But it seems that as a matter
>of logic, if 'a=b', where 'a' and 'b' are names, then necessarily 'a=b'
>
>I suggested that in this framework proper names are not being used as
>labels for objects. Rather they are functions which pick out objects
>of a certain description.
>
>I think Thony and I now agree that the question is whether names ought
>to work this way.

I'm a little concerned about this bit. Remember, one of the central points of the 2-D framework is that when S is necessary (in the usual sense), S has a necessary secondary intension. So what follows from the necessity of "a=b" is that "a=b" has a necessary SI, i.e. that 'a' and 'b' have the same SI. Nothing about PIs follows. In particular the claim that 'a' and 'b' have different PIs is completely compatible with the claim that "a=b" is necessary. (Just look at "Hesperus = Phosphorus", etc.)

Once can think of the PI as being what picks out reference in the actual world (depending on how it turns out) and the SI as picking out reference in counterfactual worlds (given that the actual world is fixed). I.e., PI gives reference across worlds considered as epistemic possibilities, SI across worlds considered as subjunctive possibilities. Given this, then in a case such as the above (with different PI, same SI), 'a' and 'b' will pick out the same referent in all counterfactual worlds (all subjunctive possibilities), as Kripke et al lead us to expect. All that follows from the difference in PIs it is epistemically possible that reference comes apart, i.e. there are epistemic possibilities such that if they turned out to be actual, reference would come apart, which seems to be correct in these cases.

Proper names are still being used as labels for objects in this framework, and they are still rigid designators. They pick out the same object across counterfactual worlds (considered as counterfactual). There's a sense in which actual-world reference is mediated by something a bit like a description (see previous note), but that's a far cry from saying that names here function like descriptions (which are nonrigid, etc, and have very different SIs).

Witness the difference between "Hesperus" (the name) and "evening star" (the description).

I had hoped that all this was clear by now, but the above suggests that maybe there is still a little confusion. Or maybe I am just misreading the above, in which case feel free to elaborate on the issue.

>> Brad tried to provide motivation for thinking that names do in fact have
>> descriptive content. One such motivation comes from considering cases in
>> which co-referring names can't be substituted *salva veritate* (such as in
>> belief contexts). He also mentioned substitution failures in "simple
>> sentences" (see Saul 1997), such as "Lois Lane kissed Superman before she
>> kissed Clark Kent".

This is interesting indirect evidence, but I don't think the case for PIs relies on this sort of evidence. The mere epistemic possibility that Superman is not Clark Kent, and the way the reference of a term varies depending on which epistemic possibility turns out to be actual, is enough to make the case. Again, these things don't imply descriptive reference across counterfactual worlds, but they do arguably imply some quasi-descriptive content across epistemic possibilities.

Incidentally it's probably a good idea not to get too fixated on issues specific to proper names, as many of those issues (especially those of reference via linguistic causal chains) aren't all that crucial to what comes later. Natural kind cases are probably more crucial.

>We then dicussed my BIV cases. From an intuition point of view we
>were evenly split about whether or not there was really any water
>after you were pulled out of the vat. Brad suggested that it may just
>be very difficult to think of this case as anything other than a
>person jumping between two possible worlds. This, of course, is
>impossible. But it does look kind of like that.

Well, it may be more palatable to look at this as a person jumping between two very different parts of one world. A bit like going from one planet to another. In this case, going from a "virtual environment" to a real environment. Different environments in one sense, but still all part of the same world (it's a big world!).

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 02:05:16 1999
Date: Fri, 19 Feb 1999 02:05:05 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: possibility?
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Anthony writes:

>I think this is probably getting somewhat off the topic, but I have a
>general worry about making sense of secondary intensions.
>
>The 2-intension of water is H₂O, given that, in this world, we take it
>that this chemical structure explains the various properties we associate
>with water. More generally, it seems that secondary intensions are
>different from the primary intension only when there is some sort of
>reductive explanation of a particular term or thing (p62).

Well, I'd say this is often the case (especially for natural kind terms), but it needn't be the case. E.g. "one meter" has different PI and SI, but no reductive explanation in the vicinity. Similar for "Hesperus". It's true that for natural kind terms, though, we very often identify it with some explanatory microstructure across worlds. So the SI of "water" goes with H₂O, of "tiger" goes with some bio-plus-DNA kind, of "heat" goes with molecular motion, etc. I don't think an actual explanation is required -- arguably the SI of "water" picked out H₂O in all worlds even before we had an explanation. But it's arguable that the SI of most such terms picks out the explanatory microstructure of the stuff, whatever that happens to be in the actual world, across all worlds.

>It is not, of course, directly observable that water is H₂O. Rather, it is
>the result of theory constructed from a variety of experimental results.
>Consider a possible world in which water is H₂O, but in which the natural
>laws are such that the elements behave radically different than they do in
>our world. Suppose, for example, that Hydrogen atoms form extremely
>unstable compounds when they are in the presence of other Hydrogen atoms.
>In this possible world, the dominant 'watery stuff' is XYZ. Thus, XYZ is
>the stuff picked out by the primary intension of water. XYZ behaves
>experimentally much like H₂O behaves in our world. But XYZ is not composed
>of two Hydrogen atoms and an Oxygen atom in this possible world.

OK, it seems plausible that the PI of "water" picks out XYZ in this world. The Kripkean may well say that the SI of "water" picks out H₂O, irrespective of its instability.

>It seems to me that the secondary intension of water cannot just be H₂O.
>It must also include a host of other information regarding the underlying
>physical theory that makes this identification meaningful. If there is a
>possible world in which XYZ behaves for the most part like H₂O, only X, Y
>and Z are not 2 H's and an O, it seems strange that there is any value in
>the secondary intension of water picking out H₂O, stuff that is, in this
>possible world, unstable.

Interesting point. The Kripkean often talks as if any H₂O will be water. They will then describe this world as a world in which water molecules are unstable. It does arguably make sense to consider subjunctive possibilities such as "water molecules might have been unstable", or "if water molecules had been unstable, life on earth would not have evolved". If so, then such worlds (considered subjunctively) are worlds with water molecules in them, and thus (presumably) worlds with water in them.

Still, I can see your intuition that thus might not really deserve to count as "water". One could similarly consider worlds where H₂O molecules make multicolored paints that people hang on walls, or are used in brick buildings, etc. Are these worlds where paintings are made of water, etc? (Even in the subjunctive SI sense?) I can at least see the counter intuition, though I go both ways.

How to accommodate that intuition? Maybe (1) Insist that the SI of water still picks out H₂O in all worlds, but say that that the molecules only count as H₂O if they are quite a bit like our H₂O molecules with current chemistry (i.e., being H₂O requires more than just having 2 H's and an O).. Or (2) identify the SI of water with something like "H₂O, as long as it has at least some of the other chemical or surface properties it has in the actual world". Or (3) go all the way to SI = "watery stuff", which would be a more radical move, letting in XYZ as water in some counterfactual worlds;

unorthodox, but some (e.g. David Lewis) have been tempted to argue that there are worlds where water is XYZ.

>If this isn't just confused, it may be that this is a reason why you say
>primary intensions are particularly important to your account.

Well, I do agree that there is an awful lot of slack in what counts as the SI of a given term. It depends some pretty debatable subjunctive intuitions about what it takes be water in a counterfactual worlds. There is also some slack in a PI, but arguably not so much -- we're somewhat clearer about what we'd count as water in different ways the actual world might turn out. This isn't the main reason I think PIs are more important than SIs (the central reason is that I think PIs tie much more closer to the epistemic notions that are at the heart of content, thought, and modality), but it may be one way of getting at the fact that SIs are most relevant in subjunctive and counterfactual contexts, about which our intuitive descriptions are pretty debatable.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 05:25:06 1999

Date: Fri, 19 Feb 1999 05:23:40 -0800

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Apriority, necessity, conceivability, possibility

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Recent discussion has been interesting, but we've drifted a little from the central topics for which the 2-D framework will be important for us. In particular, the framework is most useful for analyzing the relationship between apriority and necessity, the notion of the necessary a posteriori, the relationship between conceivability and possibility, and the issue of a priori access to modality. Of course all of these have are central to important issues in metaphysics, not least to analyzing epistemic and modal arguments in the philosophy of mind. So it's worth going over a few of these things.

Most of what's contained in this note is very important to the courset, so read it carefully.

DEFINITION OF PRIMARY AND SECONDARY INTENSIONS

I'll take a moment to officially define PIs and SIs for statements. So far we've mostly been talking about PIs and SIs for terms, but often it is whole statements that are most crucial. In effect, the PI of a statement is what you get from "composing" the PI and SI of its component terms. Note that I use "primary proposition" and "secondary proposition" for this idea in the book, but now I prefer to use "primary intension" and "secondary intension" for both.

The primary intension of a statement S is a mapping from centered worlds to truth-values. The secondary intension of S is a mapping from uncentered worlds to truth-values. They're defined as follows:

The primary intension of S maps a centered world W to the truth-value of S in W when W is considered as actual (i.e. considered as an epistemic possibility).

The secondary intension of S maps a centered world W to the truth-value of S in W when W is considered as counterfactual

(i.e. considered as a subjunctive possibility, with the actual world fixed).

We can give a couple of heuristics for evaluating the PI or SI of a statement S at a world W.

PI Heuristic: If W turns out to be actual, will S turn out to be true?

SI Heuristic: Given that the actual world is the way it is, is S true in world W?

Take W = XYZ-world, S = "water is XYZ". If the XYZ-world turns out to be actual, "water is XYZ" will turn out to be true, so the PI of S is true in W. But given that the actual world is the way it is (i.e. the H₂O-world), "water is XYZ" is false in the XYZ-world, so the PI of S is false in W. (I abstract away from centering issues here.)

(My favorite heuristics are actually a bit more streamlined than this, though maybe a bit less self-explanatory:

PI Heuristic: If W is the case, is S the case?

SI Heuristic: If W were the case, would S be the case?

This ties PIs to indicative conditionals (which basically force you to evaluate the antecedent as an epistemic possibility) and ties SIs to subjunctive conditionals (which make you treat the antecedent as some sort of "fantasy"). But I'll say more on this later.)

APRIORITY AND NECESSITY

The 2-D framework can be understood as making the following claims concerning the necessity and apriority of statements S.

NECESSITY THESIS: S is necessary iff S has a necessary secondary intension.

Here, I'm taking "necessary" in the standard Kripkean sense (i.e. "subjunctive necessity"; I'll argue later that this usage is not wholly uncontestable, but set that aside for now). S is necessary in the Kripkean sense precisely if it is true in all worlds considered counterfactually, which is just to say that its secondary intension is true in all worlds. I hope this claim is fairly straightforward, given the definition of a secondary intension.

APRIORITY THESIS (weak form): If S is a priori, S has a necessary primary intension.

The case for this is pretty straightforward. If S is a priori, then plausibly it will be true no matter how the actual world turns out, i.e. it will be true in any world considered as actual (considered as an epistemic possibility). That is, for an arbitrary world W, if W turns out to be actual, S will turn out to be true. That is, the primary intension of S is true in all centered worlds.

APRIORITY THESIS (strong form): S is a priori iff S has a necessary primary intension.

This makes the further claim that if S has a necessary PI, it is a priori. This isn't entirely straightforward to "prove". Necessary PI means that S will be true no matter how the actual world turns out,

but it is still a leap from there to the epistemic claim that we can know S to be true. Maybe S could turn out to be unknowable, for example. E.g., maybe some really complex mathematical truth could have a necessary PI but not be knowable a priori?

We'll come back to this issue of "scrutability" later. For now, I'd consider the weak form better established than the strong form, but I'd also say it's plausible that the great majority of statements with a necessary PI seem to be knowable a priori, so that exceptions to the strong form are rare if they exist.

I'm very interested to see any potential counterexamples to any of these principles (or arguments). It's not impossible that I've overlooked something here, so feel free to jig them up.

A POSTERIORI NECESSITY THESIS: S is an a posteriori necessity iff S has a necessary secondary intension and a contingent primary intension.

This follows directly from the necessity thesis and from the apriority thesis (strong form). I'd say that this is the central claim of the 2-D analysis of a posteriori necessity. It certainly seems to fit all the Kripke examples: "Hesperus is Phosphorus", "water is H₂O", etc.

Somebody might deny this thesis (presumably by denying the strong form of the apriority thesis), asserting that there is some other sort of a posteriori necessity. I think it's at least highly plausible that such a posteriori necessities would have to have their roots in considerations quite distinct from the ones Kripke raises, though.

One reason all this is useful is that it lets us jump from epistemic claims about what is and isn't a priori to modal claims about the existence of various possible worlds. E.g. if S isn't a priori, there's a possible world where its PI is false; and even if S is an a posteriori necessity, its PI will be false in some world. If this is so, it's very important to the epistemology of modality, and to the role of epistemic arguments in metaphysics.

What makes all this possible is a central property of the 2-D framework: that it has two dimensions of semantic evaluation (two intensions), but just one space of worlds. Centering aside, PIs and SIs are evaluated at exactly the same worlds. They just give different results because of the different ways the worlds are considered (as actual or as counterfactual). Thus the worlds considered as "epistemic possibilities" for PI purposes are perfectly respectable metaphysically possible worlds: witness the XYZ-world, for example. Just one world, but two different ways of looking at it. That means that when we have an a posteriori necessity, we have a genuine metaphysical possibility in which the PI of the statement is false. That metaphysical possibility may well have consequences for metaphysics.

Again, any potential counterexamples and counterarguments are welcome.

CONCEIVABILITY AND POSSIBILITY

We can also put all this in terms of the relationship between conceivability and possibility. Conceivability is an epistemic notion. There are various notions in the area, as we'll see, but perhaps the simplest is the following. (N.B. I now prefer this way of doing things to the way I do things in the book.):

S is conceivable iff S is not ruled out a priori (i.e., if it is not a priori that not-S).

If the 2-D framework is right, then one will be able to make an inference from conceivability to possibility. If S is conceivable, not-S is not a priori, so not-S has a contingent PI (by the strong apriority thesis), so there is a metaphysically possible world in which the PI of S is true. That's not to say that S itself is "metaphysically" possible in the Kripkean sense (that sense requires a world where the SI is true), but we still have a metaphysically possible world in the vicinity. Call this the

CONPOSS PRINCIPLE: If S is conceivable, there is a possible world in which the primary intension of S is true.

We'll be looking at different versions of this claim later, and examining just what it takes to support the relevant versions. But note for now how it fits the Kripkean examples. People often say that the Kripke examples show that there can be no inference from conceivability to possibility. E.g., in the relevant sense of conceivability, "water is XYZ" is conceivable, as is "Hesperus is not Phosphorus", but neither of these is possible. But on the 2-D framework, there is at least a possibility in the vicinity: a possible world in which the PI of "water is XYZ" is true, and one in which the PI of "Hesperus is not Phosphorus" is true.

Most importantly, when you conceive (in the relevant sense) of a scenario in which water is XYZ, or in which Hesperus isn't Phosphorus, you **do** conceive of a perfectly respectable world: the XYZ-world in the first case, and a world in which the evening object isn't the morning object in the second. These worlds don't satisfy the statements in question when they're considered as counterfactual, but they do when they're considered as actual. That's no surprise: after all, one is conceiving the scenarios in question as epistemic possibilities (i.e. considering them as actual), so one would expect that the worlds that correspond to this conceivability are worlds that satisfy S when they are considered as actual. So it's a little more indirect than a simple inference from S is conceivable to S is possible, but we still have conceivability as a guide to metaphysical possibility in some sense.

MODAL RATIONALISM

Note that the two-dimensional framework fits nicely with modal rationalism: the claim that we have a priori access to modality and to the space of possible worlds. Conceivability as described above is a paradigm a priori method. And when we conceive of a scenario in the relevant sense, it appears that we really are conceiving of a possible world (the XYZ-world, or the two-star world). Of course the sense in which it's an S-world is that it satisfies S's primary intension, not its secondary intension. So we have to be careful about just how we describe the worlds in question. But still, conceivability acts as a direct guide to the space of possible worlds. So the way is opened for some sort of modal rationalism.

Of course all this depends on the central claims of the 2-D framework being correct. I think there are probably a few points at which an opponent can argue (e.g. the strong apriority thesis), but it's tricky, and again they'd have to go well beyond Kripke.

PHILOSOPHY OF MIND

All this is very relevant to epistemic arguments against materialism. One response to these arguments is to deny that there is an epistemic gap between physical and phenomenal, and assert that there is an a priori entailment. I sometimes call this "type-A materialism". It's a hard row to hoe, though, because it only seems to work if one accepts a functional analysis of phenomenal concepts, and such an analysis seems very implausible.

So perhaps the most popular response is "type-B materialism": accept that there is an epistemic gap, but deny that there is an ontological gap. This doesn't have the intuitive implausibility of the type-A view, but still holds onto materialism. Typically, one will accept that there is no a priori entailment from physical to phenomenal, but will assert that there is a necessary entailment. So the entailment from physical to phenomenal is an a posteriori necessity.

This view also usually goes along with the claim that conceivability doesn't imply possibility. So zombies (for example) are said to be conceivable but not possible, so materialism is not endangered.

If the central claims of the 2-D framework are correct, this isn't so simple. The type-B materialist holds that " $P \rightarrow Q$ " is an a posteriori necessity, where P is the physical truth about the world and Q is a phenomenal truth. On the 2-D framework, it follows that " $P \rightarrow Q$ " has a contingent PI, so there is a possible world where the PI of P is true, but the PI of Q is false. Under a bit more analysis, the existence of this possible world is highly threatening to materialism.

Similarly, if zombies are conceivable, it follows that there's a possible world in which at least the PI of "there are zombies" is true. From here (under a bit of analysis) one can plausibly get to a world physically identical to ours but different overall, so that materialism comes out false.

The basic reason for all this, of course, is that deep down the Kripkean framework hasn't really broken the link between conceivability and possibility, it has just recast it a little. But we've seen that even Kripke's own results rely on the method of conceivability: conceive of a world, then make sure you describe it correctly. So it arguably supports rather than opposes a modal rationalism at the deep level. If so then the Kripkean framework gives much less support to the type-B materialist than is often supposed.

We'll be going into the full details of how this applies to materialism etc next week, so we can save a detailed analysis of the consciousness issues for then. But it's important to see how the 2-D framework might have significant metaphysical consequences, and how those rest pretty straightforwardly on the central theses of the framework.

For now, I'd be interested to see some discussion of those theses, and of the relationship between apriority and necessity vis-a-vis PIs and SIs, etc. All comments are welcome.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 06:21:34 1999
Date: Fri, 19 Feb 1999 06:20:07 -0800

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Next week
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

My visa still isn't through, and although there are indications that it shouldn't be too long now, I won't be there by next Tuesday. This means I have to rearrange things a little. Week 5 was originally supposed to be on the "tyranny of the subjunctive", on relationships between the 2-D framework and indicative and subjunctive conditionals, and on implications for the analysis of necessity. But that isn't really written up (except very briefly in the Princeton outline); I'd been planning to mostly talk about it in class. So as things stand, we will skip straight to the following topic (the original week 6) instead.

This topic involves applying the 2-D framework to epistemic and modal arguments against materialism. The readings are

TCM, Chapter 4 -- mostly pp. 131-149.
N&N, pp. 144-55.
Mind and Modality, Lecture 1 (esp. sections 5-7).

The Mind and Modality notes (on the web) have my currently preferred formulation of the 2-D argument against materialism. It's a bit crisper and tighter than the book version, though it comes to much the same thing in the end. The book version has more discursive detail, so you'll probably want to look at the two of them together (the real core of the book argument is pp. 131-36, but pp. 147-49 also has relevant material). It might also be interesting to look at Kripke's anti-materialist argument in N&N, which is similar in spirit, and compare and contrast. (N.B. The Princeton notes have some contrastive analysis that goes a bit beyond what is in the book.)

Anyway, consider these the readings for next week's discussion. I'll look forward to seeing your thoughts on them early next week.

As promised, there's also an assignment. I'd like you to take two terms that Kripke discusses, and translate his discussion into the 2-D framework. One is "yard", as discussed on p. 76; the other is "cat", as discussed on p. 122 and pp. 125-6. I'd like you to take every significant part of Kripke's discussion here and translate it into the 2-D framework. Characterize roughly what the PI and the SI of these terms look like, at least given Kripke's intuitions. In places where Kripke considers a hypothetical scenario, you should roughly specify the world, say whether it is being considered as actual or as counterfactual, note what the referent of the relevant term or statement is in that world (considered the relevant way, and according to Kripke's stated intuitions), and say what the upshot is the for relevant PI or SI. Try to translate Kripke's more general points in these passages into the framework too, if you can. If you're so inclined, you can say whether you agree or disagree with Kripke's specific and general analyses and why, though that isn't compulsory.

This shouldn't be much work (the passages are pretty short). This will be due by next Tuesday at noon, Arizona time. Don't be late. E-mail it directly to me (not to the mailing list). It should be your own work.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 18 08:10:40 1999
Date: Thu, 18 Feb 1999 09:09:39 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

>
A quick comment on Dave's comment on my comment on Josh:

[For Dave's comments, and some of my original comments, see below.]

The point I was making about glumpy stuff was this. In Josh's scenario, I'm told (and believe) that I've just left a Vat. Presumably I will also believe that none of the stuff behind my past experiences is the same sort of stuff as the stuff behind my current experiences. In this situation, I want to suggest that we shouldn't reidentify my in-Vat <"water"> concept with my post-Vat water concept, *even if we're picking out concepts in terms of their primary intensions.* It is true that both in the vat and out of the vat, I have a concept that is closely connected with my phenomenal concept <watery stuff>, and it's a natural kind concept too. Call these two concepts 'Water1' and 'water2' (These are just two names, I'm not yet supposing that these are two distinct concepts). When, out of the Vat, I use water2, I don't think of it as applying to the same sort of stuff that water1 applied to. And that, it seems to me, is enough to say that its primary intension is different.

Here's an analogy. Suppose I have two concepts, <glumph> and <water>. I think that glumph and water are distinct natural kinds. Both act in the same (watery) way, indeed, I can't tell them apart. Almost nobody can tell them apart. (Perhaps nobody now living can tell them apart. It's just part of our mythology that there were people - the shabby pedagogues - who could tell them apart.) But I know that the stuff around here is glumph. Still, as far as I'm concerned, I interact with glumph pretty much the same way I would with water. Now, suppose that I move to an area where the watery stuff is not glumph but water. *If I know that this is the case*, I will use a different concept, my <water> concept rather than my <glumph concept>, to refer to the watery stuff around me.

I want to say that this case is precisely analogous to Josh's BIV case in which he believes that he was a BIV. It may look like he is still using his water concept, but he's not. He's using a concept that has the same surface structure, but there's more to PI than surface structure. At least for Natural Kinds, one has to think that one is talking about the same sort of stuff in order for "two" primary intensions to be the same.

Well, maybe.

> There's also the special case where I'm told I've shifted, and where I
> make a new stipulation about my concepts so my beliefs don't come out
> useless. That would plausibly involve a "different concept" in a
> reasonably strong sense, but that also seems to be something of a
> special case.
>
> Tim raises some worries about concept identity:
>
> >But should we think that there is any sense in which VAT-Josh

> >and post-Vat Josh have a concept, i.e. <water> that can be tracked from
> >the Vat to after the Vat? I wonder. Both in the Vat and after the vat Josh
> >has a concept that is related to <watery>, but that doesn't seem
> >sufficient to justify the claim that he has a single concept that survives
> >the experience from Vat to actual world, that we can reidentify as the
> >same concept. After all, both my concept of a certain chemical substance
> >(say, glump - suppose that glup is watery) and my concept <water> are
> >connected to <watery stuff>, but that doesn't give me any incentive to
> >identify <glump> with <water>.

>
> I presume you're talking about the vatted concept and the
> long-post-vat concept. In favor of "same concept" is that they have
> the same PI, and a continuous history of use between them. Against
> "same concept" is that they have a different referent and SI. What
> answer one gives will partly depend on whether one individuates
> concepts by PI, SI, referent, or history of use. I think this is
> somewhat terminological. I have some temptation toward "same
> concept", given the centrality of PIs, but then the difference in
> referent and SI is pretty significant, too. As for water/glump, I'm
> not sure how to imagine the case -- do they pick out the same
> referent, or not? If the same referent, presumably we have same SI,
> different PI, so the reverse of the above. If a different referent,
> presumably we have different PI and SI, which is worse (presumably
> even if the PI of "glump" involves "watery", it will be in a different
> way, as it isn't a priori that water is glump). Either way, one
> relevant difference is that the vat case involves sameness of PI and
> continuity of use, whereas this case doesn't.

>
> >To sum: supposing the individualistic spin on water, in which the primary
> >intension is <whatever has been causing *my* watery experiences> we (and
> >Josh) might still be tempted to identify his in-Vat watery concept with
> >his post-Vat watery concept, even though it refers to a different stuff,
> >because many of the inferential connections that he learnt while in the
> >vat still hold true in the actual world. If post-Vat Josh thinks of his
> >envatted experiences as including water thoughts, then it is clear that he
> >must have a water concept that we can trace from the Vat to the real
> >world.

>
> I think I agree with this, more or less. There are actually two
> issues: (1) sameness of concept pre- and post-vat, and (2) whether the
> concept picks out the actual stuff. That the "inferential connections
> still hold true in the actual world" seems to be getting at (2) rather
> than (1). In my terms, I'd put it by saying that as long as the
> actual world isn't *too* different from the vat world, actual watery
> stuff can come to satisfy the PI of my "water" concept (though
> possibly with a time lag). I'm inclined to think that (1) doesn't
> turn so much on the actual facts, but more on facts about the
> subject's psychological structure (modulo issues about individuating
> concepts by referent). That seems to fit with your last sentence
> above.

>
>
> (3) Tim raises the case where I'm shifted to Twin Earth in my sleep.
> As I said in the last message, I think Tim is right that my "water"
> won't initially refer to XYZ, though it may after a couple of years.
> I think the moral of that is that the PI of "water" anchors one to
> one's past environment, as Tim suggests. Basically because it picks
> out (at least in part) the stuff that one has causally interacted with
> in the past. In fact I think that for many of these cases, the
> present environment is mostly irrelevant (except to determining
> whether one's beliefs are now true or false, etc). The future

> environment even more so, though maybe one could come up with cases
> where it matters.
>
> All thoughts and reactions are welcome!
>
> --Dave.
>

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 19:38:16 1999
Date: Fri, 19 Feb 1999 19:38:01 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Tim suggests that maybe the BIV pre- and post-emergence should be regarded as having two sets of concepts, stemming from the fact that he knows he has changed environments. I think this isn't implausible. It's related to what I said earlier about the BIV being able to make a new "stipulation" about its concepts once it knows it's environment has changed so much (or once it knows it was previously so deceived, depending on how you do things).

In the case where the ex-BIV doesn't know about the change, presumably this won't work -- it will keep using the old concepts, and for a good while, at least, it will make a lot of claims that are strictly false (e.g., "there's water around here"). Even if the ex-BIV knows about the change, presumably it doesn't need to make the stipulation -- it might correctly say "there's no water around here, just stuff which looks a lot like the stuff I knew as water". But that would probably lead to inconvenience and communication problems for the ex-BIV, who might prefer truth and communication, so it might decide to adopt the conceptual system of his new community wholesale. Of course the new community will be using the same words (or same sounds) as the old community, but that's fairly incidental -- he's really deciding to use new concepts. So maybe his old "water" concept had as PI "the watery-looking stuff I was causally connected to", and his new one has PI "the watery stuff that's around me recently", or "what people call 'water' around here", or something like that.

That does seem to be a special case, stemming from (a) the fact that he knows about the change, and (b) the fact that due to some sort of coincidence and conspiracy, his new environment is enough like his old environment that his old words "sort of" work. If one had (b) without (a), one presumably wouldn't get this sort of new concept. And if one had (a) without (b), maybe he'd adopt a new set of concepts, but presumably they wouldn't have such a close relationship to the old concepts (quite different semantics, probably different sounds). But given the special circumstances of (a) and (b), I think Tim's analysis is pretty reasonable.

>When, out of
>the Vat, I use water₂, I don't think of it as applying to the same sort of
>stuff that water₁ applied to. And that, it seems to me, is enough to say
>that its primary intension is different.

That's a nice point. The fact that a person knows that a new concept doesn't apply to what the old concept applies to is a good indication that it's a different concept. It's a bit like the a posteriority of "a=b" implying that 'a' and 'b' have different PIs. Here, he can still use his old concept if he wants, and then he'll hold that "water₁ = water₂" is false, and so certainly not a priori. So one has different concepts. Unlike the case where he doesn't know about the change, and can't entertain the false identity.

The glumph/water case is a nice way of making the point too. The main disanalogy is that here one had both concepts beforehand, but otherwise things seem similar. Again, the fact that "water = glumph" can be false, or even its non-apriority, seems to be a sign of two different PIs and two different concepts.

>At least for Natural Kinds, one has to think that one is talking about the
>same sort of stuff in order for "two" primary intensions to be the same.

That has a certain plausibility. Certainly if one has two concepts 'a' and 'b' and holds that "a=b" is false, that suggests two different PIs for the two concepts.

A slight complication, though: There may be cases where I don't so much have two concepts, but hold that "what I now refer to as 'a' is different from what I used to refer to as 'a'" -- e.g., take 'a' = 'here', or 'the president'. Presumably this is compatible with 'a' having the same PI, just a different referent due to the change in location. That's due to a special property of indexicals -- they're not supposed to have a stable referent. Can this happen for natural kind terms? Well, I suppose it could happen if one goes to twin earth, keeps using one's old "water" concept, the reference gradually shifts to XYZ, and only then does one find out about the change. That might be a case where the PI has stayed constant. So maybe one has to qualify your thesis above.

What exactly is the different between the ex-BIV shift and these two cases? Something to do with the fact that natural kind referents can't shift *quickly*, given just the one concept. In the ex-BIV case, the old PI should pick out the old stuff just after moving, not the new stuff. So if one has the possibility of a different referent so quickly, that indicates a new PI and a new concept. Of course that constraint doesn't imply to standard indexicals, whose reference can shift in a flash.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Feb 21 19:55:57 1999
Date: Sun, 21 Feb 1999 20:55:04 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Angela brought to my attention a case we were discussing a while back

involving quantum mechanics that may have relevance for evaluating the strong apriority thesis; specifically the conditional from "S has a necessary primary intension" to "S is a priori". Consider some quantum mechanical superposed state (of some arbitrary photon impinging on a half-silvered mirror, say). The state is described by the mathematics of Shrodinger's "U" equation, or the linear complex number equations governing the evolution of the photon as it superposes into different possible states. Now, since this is entirely mathematical it ought to have the same status as any "complex mathematical truth"; hence, if the latter has a necessary primary intension so will the former. The actual position of the particle, however (which is what the equations is tracking) is not knowable apriori. After the "reduction" or collapse of the wave function, some distribution of probabilities will tell us where (in what final state) the photon will be, but even now there is no apriori truth--accessible by us anyway--of what the final state of the proton will be in. Worse, there is no way of knowing even while the unitary evolution is underway, although "the position of this particle" is presumably the primary intension of the equation. If we change this to "the possible positions of this particle" then among most construals of the U-evolution (which is supposed to be entirely deterministic prior to collapse) we are not getting the right primary intension. So it seems in this case (although I admit this is the subject of much debate) that, if an equation of this sort has a primary intension (and why shouldn't it, since it is just some mathematical statement), then the real content of it is not knowable apriori. There is a big debate over whether to treat the collapse or state vector reduction as real or imaginary, and also whether to consider the linear superposition of subatomic phenomena as descriptions of real superpositions or just incomplete descriptions of particle positions (I suppose these are slightly different issues), but the point is that in as much as something like a Schrodinger equation has a primary intension, it is about the positions of particles, and this is not knowable apriori in a very clear cut way, in this case.

One other one, take as a formal system the smartest mathematician, construct his Godel number. There is a truth not knowable to any of us but knowable in principle apriori.

Erik L.

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Mon Feb 22 03:52:31 1999
Date: Mon, 22 Feb 1999 03:51:31 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Erik raises a couple of intriguing potential counterexamples to the strong apriority thesis (the claim that S has a necessary PI iff S is a priori). I'll take the second first, as I think I understand it a

bit better.

>One other one, take as a formal system the smartest mathematician,
>construct his Godel number. There is a truth not knowable to any of us
>but knowable in principle apriori.

Right. On the assumption that each of us is a finite computational system, it follows by a natural chain of reasoning from Godel's theorem that there are mathematical truths that none of us can know to be true. If G is such a statement for the smartest of us (or for the whole community), then presumably G is a necessary truth, and has a necessary PI (there's no difference between PI and SI for mathematical statements). But none of us can know it a priori. Counterexample?

This is one of the trickiest cases for the strong apriority thesis to handle. But I think it can be handled by noting (as you note in your last sentence implicitly) that we are using an idealized notion of the a priori. S is a priori if it is knowable a priori, and in considering what is knowable a priori we abstract away from our contingent cognitive limitations. On this way of doing things, it's plausible that G comes out to be a priori. It's true that none of us can know it, but that's just because of our cognitive limitations. There are presumably smarter beings that can come to know it a priori. So G is a priori.

[Semi-technical note (don't worry if you don't follow this): To justify this, one can note a result of Kleene's (I believe, though I may be misrecalling the details) to the effect that any arithmetical statement that is undecidable within Peano arithmetic (PA) is decidable in some extension of PA, where the extensions are obtained by repeated "Godelizing" (adding a Godel sentence), onward through the ordinal hierarchy. The trouble with us is that we can't Godelize forever -- being finite systems, we eventually lose track of ordinal counting (which is itself not recursively systematizable). But for any undecidable statement of arithmetic, *some* amount of repeated Godelization will decide it. So all we need is a being smarter than us who can count further through the ordinals. Presumably there's no obstacle to the possibility of such a being; our own specific limits here seem to be contingent cognitive limitations. (Shaughan, feel free to correct or expand on this!)]

This is some heavy machinery to bring out, and arguably there are other strategies. Sometimes I'm even tempted to appeal to the possibility of a being who can go through all the integers at once to determine whether "for all n, P(n)" is true, on the theory that our inability to do this is a contingent cognitive limitation, but that might make one a little queasy. There are also questions about what to say for higher set theory and the like. We'll be coming back to these matters a bit later on, but the important thing to note is that for the strong apriority thesis to be plausible, we have to appeal to a notion of the a priori that idealizes away from our contingent limitations.

Incidentally I discuss the relevance of this sort of analogy to the mind-body discussion in the book briefly around pp. 138-40. The way I prefer to do things now, in terms of positive and negative conceivability (coming up in two or three weeks), it turns out fortunately that one doesn't need to refute this sort of example at least to make the case re consciousness. A thesis slightly weaker than the strong apriority thesis can do the job for the anti-materialist, and this sort of consideration gets factored out.

It still raises independently important issues for modal rationalism, though.

Feel free to follow up here, as there are plenty of interesting issues in the vicinity.

>Consider some quantum
>mechanical superposed state (of some arbitrary photon impinging on a
>half-silvered mirror, say). The state is described by the mathematics of
>Shrodinger's "U" equation, or the linear complex number equations
>governing the evolution of the photon as it superposes into different
>possible states. Now, since this is entirely mathematical it ought to
>have the same status as any "complex mathematical truth"; hence, if the
>latter has a necessary primary intension so will the former.

I'm not sure exactly what the sentence is whose PI we are talking about. But in any case I think I might get off the bus here. I don't think the Schrodinger equation has quite the same status as a mathematical truth. As it's being used here, it's a **physical** truth, and as such is both a posteriori and has a contingent PI. Of course the mathematical truth that a certain differential equation has such-and-such solutions is a priori (and has a necessary PI); but here we are concerned with the claim that these solutions correspond to the way things are in physical reality, and that's a different claim altogether.

>The actual
>position of the particle, however (which is what the equations is
>tracking) is not knowable apriori. After the "reduction" or collapse of
>the wave function, some distribution of probabilities will tell us where
>(in what final state) the photon will be, but even now there is no apriori
>truth--accessible by us anyway--of what the final state of the proton will
>be in. Worse, there is no way of knowing even while the unitary evolution
>is underway, although "the position of this particle" is presumably the
>primary intension of the equation. If we change this to "the possible
>positions of this particle" then among most construals of the U-evolution
>(which is supposed to be entirely deterministic prior to collapse) we are
>not getting the right primary intension. So it seems in this case
>(although I admit this is the subject of much debate) that, if an equation
>of this sort has a primary intension (and why shouldn't it, since it is
>just some mathematical statement), then the real content of it is not
>knowable apriori.

Well, let's separate a few things. (1) The mathematical fact that the Schrodinger equation has certain solutions. That's a priori and has a necessary PI. (2) The physical fact that the Schrodinger equation describes physical reality. That's a posteriori and has a contingent PI. (3) The physical fact that a certain particle has a certain position. That's highly a posteriori and has a contingent PI.

(More technicalities coming up.)

I think what you may really be talking about, though, is the conditional from initial state of the wave function plus Schrodinger equation (as a physical law) to final state of the particle. The status of this depends on just which interpretation of QM we are using, as the claim the Schrodinger equation is taken to make about physical reality will vary with the interpretation. On the Everett interpretation, I suppose the conditional will be a priori, as the equation is the only dynamics; and indeed, given knowledge of an initial state and of the truth of the interpretation, one can know the (superposed) final state just fine. Something similar arguably goes

for the Bohm interpretation. As long as we are given full enough information about wave plus particles in the antecedent, then we can come to know the determinate final state no problem. In a "collapse" interpretation, on the other hand, the conditional will never be a priori, as such interpretations have a further nondeterministic part of the dynamics, so we'll at best be able to figure out a probability distribution. Anyway, it seems to me that in each of these cases, a priority and necessity of PI will line up.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Feb 22 10:16:12 1999
Date: Mon, 22 Feb 1999 11:14:44 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Re: 2D questions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Dave,
good point, I was using the "2-D" interpretation of QM with the U equation governing particle positions prior to "collapse" and the probability distributions after the collapse. On this view, the conditional wouldn't be apriori. One further comment, the superbeing you mentioned that can run through all integers to determine "for all n, P(n)" could not be formalizable, by Godel's result. If it was formalizable (and consistent), then it would get stuck too. If such a being could exist, it would be an "oracle" about which we could have no computational knowledge (even in principle).

Erik

On Mon, 22 Feb 1999, David
Chalmers wrote:

> Erik raises a couple of intriguing potential counterexamples to the
> strong apriority thesis (the claim that S has a necessary PI iff S is
> a priori). I'll take the second first, as I think I understand it a
> bit better.
>
> >One other one, take as a formal system the smartest mathematician,
> >construct his Godel number. There is a truth not knowable to any of us
> >but knowable in principle apriori.
>
> Right. On the assumption that each of us is a finite computational
> system, it follows by a natural chain of reasoning from Godel's
> theorem that there are mathematical truths that none of us can know to
> be true. If G is such a statement for the smartest of us (or for the
> whole community), then presumably G is a necessary truth, and has a
> necessary PI (there's no difference between PI and SI for mathematical
> statements). But none of us can know it a priori. Counterexample?
>
> This is one of the trickiest cases for the strong apriority thesis to
> handle. But I think it can be handled by noting (as you note in your
> last sentence implicitly) that we are using an idealized notion of the
> a priori. S is a priori if it is knowable a priori, and in
> considering what is knowable a priori we abstract away from our
> contingent cognitive limitations. On this way of doing things, it's
> plausible that G comes out to be a priori. It's true that none of us
> can know it, but that's just because of our cognitive limitations.

> There are presumably smarter beings that can come to know it a priori.
> So G is a priori.
>
> [Semi-technical note (don't worry if you don't follow this): To
> justify this, one can note a result of Kleene's (I believe, though I
> may be misrecalling the details) to the effect that any arithmetical
> statement that is undecidable within Peano arithmetic (PA) is
> decidable in some extension of PA, where the extensions are obtained
> by repeated "Godelizing" (adding a Godel sentence), onward through the
> ordinal hierarchy. The trouble with us is that we can't Godelize
> forever -- being finite systems, we eventually lose track of ordinal
> counting (which is itself not recursively systematizable). But for
> any undecidable statement of arithmetic, *some* amount of repeated
> Godelization will decide it. So all we need is a being smarter than
> us who can count further through the ordinals. Presumably there's no
> obstacle to the possibility of such a being; our own specific limits
> here seem to be contingent cognitive limitations. (Shaughan, feel
> free to correct or expand on this!)]
>
> This is some heavy machinery to bring out, and arguably there are
> other strategies. Sometimes I'm even tempted to appeal to the
> possibility of a being who can go through all the integers at once to
> determine whether "for all n, P(n)" is true, on the theory that our
> inability to do this is a contingent cognitive limitation, but that
> might make one a little queasy. There are also questions about what
> to say for higher set theory and the like. We'll be coming back to
> these matters a bit later on, but the important thing to note is that
> for the strong apriority thesis to be plausible, we have to appeal to
> a notion of the a priori that idealizes away from our contingent
> limitations.
>
> Incidentally I discuss the relevance of this sort of analogy to the
> mind-body discussion in the book briefly around pp. 138-40. The way I
> prefer to do things now, in terms of positive and negative
> conceivability (coming up in two or three weeks), it turns out
> fortunately that one doesn't need to refute this sort of example at
> least to make the case re consciousness. A thesis slightly weaker
> than the strong apriority thesis can do the job for the
> anti-materialist, and this sort of consideration gets factored out.
> It still raises independently important issues for modal rationalism,
> though.
>
> Feel free to follow up here, as there are plenty of interesting issues
> in the vicinity.
>
> >Consider some quantum
> >mechanical superposed state (of some arbitrary photon impinging on a
> >half-silvered mirror, say). The state is described by the mathematics of
> >Shrodinger's "U" equation, or the linear complex number equations
> >governing the evolution of the photon as it superposes into different
> >possible states. Now, since this is entirely mathematical it ought to
> >have the same status as any "complex mathematical truth"; hence, if the
> >latter has a necessary primary intension so will the former.
>
> I'm not sure exactly what the sentence is whose PI we are talking
> about. But in any case I think I might get off the bus here. I don't
> think the Schrodinger equation has quite the same status as a
> mathematical truth. As it's being used here, it's a *physical* truth,
> and as such is both a posteriori and has a contingent PI. Of course
> the mathematical truth that a certain differential equation has
> such-and-such solutions is a priori (and has a necessary PI); but here

> we are concerned with the claim that these solutions correspond to the
> way things are in physical reality, and that's a different claim
> altogether.
>
> >The actual
> >position of the particle, however (which is what the equations is
> >tracking) is not knowable apriori. After the "reduction" or collapse of
> >the wave function, some distribution of probabilities will tell us where
> >(in what final state) the photon will be, but even now there is no apriori
> >truth--accessible by us anyway--of what the final state of the proton will
> >be in. Worse, there is no way of knowing even while the unitary evolution
> >is underway, although "the position of this particle" is presumably the
> >primary intension of the equation. If we change this to "the possible
> >positions of this particle" then among most construals of the U-evolution
> >(which is supposed to be entirely deterministic prior to collapse) we are
> >not getting the right primary intension. So it seems in this case
> >(although I admit this is the subject of much debate) that, if an equation
> >of this sort has a primary intension (and why shouldn't it, since it is
> >just some mathematical statement), then the real content of it is not
> >knowable apriori.
>
> Well, let's separate a few things. (1) The mathematical fact that the
> Schrodinger equation has certain solutions. That's a priori and has a
> necessary PI. (2) The physical fact that the Schrodinger equation
> describes physical reality. That's a posteriori and has a contingent
> PI. (3) The physical fact that a certain particle has a certain
> position. That's highly a posteriori and has a contingent PI.
>
> (More technicalities coming up.)
>
> I think what you may really be talking about, though, is the
> conditional from initial state of the wave function plus Schrodinger
> equation (as a physical law) to final state of the particle. The
> status of this depends on just which interpretation of QM we are
> using, as the claim the Schrodinger equation is taken to make about
> physical reality will vary with the interpretation. On the Everett
> interpretation, I suppose the conditional will be a priori, as the
> equation is the only dynamics; and indeed, given knowledge of an
> initial state and of the truth of the interpretation, one can know the
> (superposed) final state just fine. Something similar arguably goes
> for the Bohm interpretation. As long as we are given full enough
> information about wave plus particles in the antecedent, then we can
> come to know the determinate final state no problem. In a "collapse"
> interpretation, on the other hand, the conditional will never be a
> priori, as such interpretations have a further nondeterministic part
> of the dynamics, so we'll at best be able to figure out a probability
> distribution. Anyway, it seems to me that in each of these cases, a
> priority and necessity of PI will line up.
>
> --Dave.
>

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 06:21:34 1999
Date: Fri, 19 Feb 1999 06:20:07 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Next week
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

My visa still isn't through, and although there are indications that it shouldn't be too long now, I won't be there by next Tuesday. This means I have to rearrange things a little. Week 5 was originally supposed to be on the "tyranny of the subjunctive", on relationships between the 2-D framework and indicative and subjunctive conditionals, and on implications for the analysis of necessity. But that isn't really written up (except very briefly in the Princeton outline); I'd been planning to mostly talk about it in class. So as things stand, we will skip straight to the following topic (the original week 6) instead.

This topic involves applying the 2-D framework to epistemic and modal arguments against materialism. The readings are

TCM, Chapter 4 -- mostly pp. 131-149.
N&N, pp. 144-55.
Mind and Modality, Lecture 1 (esp. sections 5-7).

The Mind and Modality notes (on the web) have my currently preferred formulation of the 2-D argument against materialism. It's a bit crisper and tighter than the book version, though it comes to much the same thing in the end. The book version has more discursive detail, so you'll probably want to look at the two of them together (the real core of the book argument is pp. 131-36, but pp. 147-49 also has relevant material). It might also be interesting to look at Kripke's anti-materialist argument in N&N, which is similar in spirit, and compare and contrast. (N.B. The Princeton notes have some contrastive analysis that goes a bit beyond what is in the book.)

Anyway, consider these the readings for next week's discussion. I'll look forward to seeing your thoughts on them early next week.

As promised, there's also an assignment. I'd like you to take two terms that Kripke discusses, and translate his discussion into the 2-D framework. One is "yard", as discussed on p. 76; the other is "cat", as discussed on p. 122 and pp. 125-6. I'd like you to take every significant part of Kripke's discussion here and translate it into the 2-D framework. Characterize roughly what the PI and the SI of these terms look like, at least given Kripke's intuitions. In places where Kripke considers a hypothetical scenario, you should roughly specify the world, say whether it is being considered as actual or as counterfactual, note what the referent of the relevant term or statement is in that world (considered the relevant way, and according to Kripke's stated intuitions), and say what the upshot is the for relevant PI or SI. Try to translate Kripke's more general points in these passages into the framework too, if you can. If you're so inclined, you can say whether you agree or disagree with Kripke's specific and general analyses and why, though that isn't compulsory.

This shouldn't be much work (the passages are pretty short). This

will be due by next Tuesday at noon, Arizona time. Don't be late.
E-mail it directly to me (not to the mailing list). It should be your own work.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 24 01:33:12 1999
Date: Wed, 24 Feb 1999 01:32:52 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: assignments
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I've been over the assignments and made comments on them. I'll send them all back individually. But here are some general comments.
First I'll give something like the answers I would have given.

"yard":

PI: picks out the distance from Henry's nose to finger in a given world.
SI: picks out this specific length (the actual Henry length) in all worlds.

Kripke's discussion: The PI of 'yard' picks out the Henry length in a given world. Even so, it's not necessary that the Henry length is a yard.

To see this, take a world where Henry's arm was foreshortened by an accident. Considering that world as counterfactual, it's a world where Henry's arm is less than a yard. I.e., "the Henry length is a yard" is false in that world considered as counterfactual. I.e., the SI of "the Henry length is a yard" is false in that world. So "the Henry length is a yard" is not necessary.

This isn't because the PI of "yard" is a cluster. The point works even when the PI is a non-cluster (as the PI used here shows). Rather it arises because the SI of "yard" is different from the PI, picking out the actual length at all worlds.

(Something else one can say, though Kripke doesn't say it explicitly: It's a priori that the Henry length is a yard. If we take the foreshortened world and consider it as actual, "the Henry length is a yard" comes out true. I.e. the PI of "the Henry length is a yard" is true in that world. Similarly for any world. So "the Henry length is a yard" is a priori. This reflects the fact that at all worlds, the PI of "yard" picks out the Henry length in that world.)

"cats":

PI: picks out "that kind of thing", i.e. the things I'm acquainted with as 'cats' or some such, at a given centered world.

SI: picks out a specific biological kind (DNA, evolution history, etc?) at all worlds.

p. 122: "Cats are animals" is not a priori. To see this, take a world where the catlike things are demons, and consider it as actual. Considered this way, we'll say that cats are demons and are not

animals. I.e., "cats are animals" is false in this world considered as actual. I.e., the PI of "cats are animals" is false in this world. So "cats are animals" is not a priori

This reflects the fact that the PI of "cats" doesn't pick out animals in all worlds. In fact it doesn't pick out any qualitative dictionary definition (for any definition D, "cats are D" isn't a priori), but rather picks out "that sort of thing".

p. 126: Given that cats are in fact animals, then "cats are animals" is necessary. To see this, take the world where catlike things are demons, and consider it as counterfactual. Considered this way, it's a world where the demons aren't cats. So "cats are demons" is false in that world considered as counterfactual. Similarly for any world in which catlike things aren't animals. Considering such a world as counterfactual, we'll always say the nonanimals aren't cats. So "cats are animals" will be true in all such worlds considered as counterfactual. So the SI of "cats are animals" is true in all worlds. So "cats are animals" is necessary. The SI of "cat" picks out animals in all worlds.

(Kripke also reiterates the PI point from p. 122: if we *discovered* that catlike things are demons, we'd say cats aren't animals. I.e. considering the catlike demon world as actual, "cats are animals" comes out false, i.e. "cats are animals" isn't a priori, and the PI of "cat" can pick out nonanimals. But here, we are considering the world as counterfactual, in order to determine what is necessary.)

Similarly even for a world in which catlike things are animals with reptile internal structure. Considering such a world as counterfactual, we'll say it's a world where the catlike reptiles are fool's cats, not cats. So the SI of "cats" doesn't pick out the reptiles in that world. So the SI of "cats" picks out animals with a specific internal structure across worlds, and so it is necessary that cats have that specific internal structure.

Of course I didn't expect precise replicas of this, but I was looking for the central points here to be made. Mostly people did a reasonably good job, showing at least a decent working knowledge of the framework. Often people could have been a bit more complete and explicit in the discussion, though. And people should work on being as precise and rigorous with the language as they can. A few points that came up repeatedly:

(1) Quite a few people didn't distinguish the epistemic discussion on p. 122 from the modal discussion on p. 126 as carefully as they could have. Often people concentrated on the modal point (re rigid designation and the SI of "cat"), without going into Kripke's argument on the epistemic point (re considering the demon world as actual, and consequences for a priority and for the PI of "cat"). It's important to distinguish the epistemic from the modal point, and recognize that what's going on on p. 122 is really quite different from what's going on on p. 126.

(2) Most people didn't address the point about clusters, and the reptile scenario. That's forgivable, as they were minor points even in context.

(3) Often people didn't quite follow my instructions re the consideration of hypothetical scenarios. In particular, it's

important to note, when a scenario is being considered, whether it is being considered as actual or as counterfactual. Then one can note what results one gets when the scenario is considered this way, and what follows regarding the PI/SI of the concept, and for apriority/necessity. (In general, scenarios considered as actual have consequences for PI and for a priority, on the epistemic side; scenarios considered as counterfactual have consequences for SI and for necessity, on the modal side.) Being careful about this makes it a lot easier to disentangle the issues in (1) above.

Note that Kripke's arguments almost always move from the specific to the general. He considers specific scenarios (as actual or as counterfactual), makes a considered judgment about how to describe them, and draws consequences about the way our terms refer across worlds (in effect about PIs and SIs), and draws conclusions about apriority and necessity. At least, that's the way the argument is presented: it's consideration of scenarios that establishes or supports the conclusions re a priority, necessity, and the like. That's a pretty common order of precedence in this area: one proceeds from a judgment about a scenario or scenarios (considered as actual or as counterfactual) to conclusions about PI/SI, apriority, necessity, etc, rather than vice versa.

(4) A common looseness of language was to move too easily between intensions and extensions. E.g. people would say that the PI is the Henry length, or that the SI is the specific length, or that the SI picks out the actual PI in all worlds, and so on. Remember, a PI is a function, not a length. So one can say things like: the PI *picks out* a length in a given world; or, the *value* of the PI at a given world is a length. But it's best not to say that the PI itself is a length.

For the phrases above, one could say e.g. that the PI picks out the Henry length in a given world, or that it picks out the actual Henry length in the actual world. Or one could just talk about the actual referent, instead of the PI, saying that the actual referent is the actual Henry length. For the second phrase, one could say that the SI picks out the specific length in all worlds. For the third, one could say that the SI picks out the actual referent in all worlds, or that it picks out the actual value of the PI at all worlds, or just that it picks out the actual Henry length at all worlds.

There are various ways to do things, but it's important for many purposes not to mix up intensions and extensions too easily. The central thing is that an intension is a function from worlds to extensions. (Where extensions are things like objects, people, species, lengths, properties, etc.) The intension is not an extension, but it picks out an extension at any given world. I don't think anyone is really confused about this, but it's good to keep the language straight.

(5) Another point about the language concerns just how we talk about evaluating statements at worlds. Given the ambiguity between PI and SI evaluation (i.e. between considering as actual and as counterfactual), one has to be careful about this. It's easy to be unclear by not specifying the sort of evaluation, or in other ways. Here are the sorts of locutions that are reasonable, where W is a world, and "A is B" is a statement.

For PI evaluation (or considering as actual), the following locutions are probably best:

In W considered as actual, "A is B" is true.
The primary intension of "A is B" is true in W.

The two above are probably the most precise, but one can also equivalently say the following without too much inaccuracy:

"A is B" is true in W considered as actual.
Considering W as actual, it's a scenario where A is B.
Considering W as an epistemic possibility, it's an epistemic possibility in which A is B.
If W turns out to be actual, it will turn out that A is B.
If W is actual, A is B.

Exactly the same goes if you substitute "counterfactual" for "actual" and "SI" for "PI" in the above, except for the last two, which don't have direct analogs. (Maybe a subjunctive such as "If W were the case, A would be B"; more on subjunctives and indicatives later.)

Something similar goes for evaluating terms rather than sentences. Here you can say things like:

In W considered as actual, "A" refers to X.
The primary intension of "A" refers to X in W.
If W is actual, "A" refers to X.

[N.B. "Picks out" is as good as "refers to" here.]

Things one *shouldn't* say:

The PI of "A is B" is true in W considered as actual. (Redundant!)
When W is considered as actual, the PI of "A" picks out X. (Redundant!)
If W is actual, the PI of "A" picks out X (Redundant!)
If W is actual, X is the PI of "A". (Intension/extension mixup.)
The PI of "A" is actually X. (Intension/extension mixup.)

One also has to be careful with things like:

In W, "A is B" is true.
In W, A is B.

When we do this without a PI/SI indicator, or without an "actual" or "counterfactual", the usual practice is that we are considering W as counterfactual, so it is SI evaluation that's relevant. At least that is the standard practice in Kripke and in most contemporary philosophy. But in the context of the 2-D framework, it can be a little confusing and potentially ambiguous, so it's best to give some sort of marker where possible.

Anyway, as I said, people did pretty well. More precision and such (very important in this area!) should come with practice and experience. I'll probably set another assignment like this fairly soon, to give more practice at the framework. Maybe including an example or two not in Kripke, for you to give your verdicts on the PI and SI, etc.

--Dave.

P.S. Now that the assignment is out of the way, I'm looking forward to everyone's comments on the week 5 issues as soon as possible.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 12 14:58:13 1999

Date: Fri, 12 Mar 1999 15:54:59 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Rachael J Parkinson <rachaelp@U.ARIZONA.EDU>
Subject: Re: your mail
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Like Angela, I am confused about what the secondary intension points to in the case of the forty-second president of the United States. To elaborate on a point that Brad made about Deep Throat in our last meeting, I can think of a few cases where rigid designators don't seem to work the way Kripke intends. For example, Mark Twain; it would seem, given in our actual world that Mark Twain is Samuel Clemmons, Mark Twain would be Samuel Clemmons in all counterfactual worlds (I guess Samuel Clemmons isn't really a secondary intension?). But I can easily imagine a world where Mark Twain was Joe Schmo. Likewise, I can imagine a world where Marilyn Monroe was not Norma Jean and Aristotle was a woman. I guess this is possible because I am assuming that the primary intension of those names apply to Mark Twain, the author; Marilyn Monroe, the actress; and Aristotle, the philosopher.

But Kripke is intent on showing that names are not descriptive. The name, Marilyn Monroe, refers to, roughly, that woman that we call Marilyn Monroe. I think that Kripke can make a good case in respect to Aristotle (because Aristotle refers to the man and not to the Greek philosopher etc. it makes sense to say that Aristotle could have been a blacksmith.) I think that it does not make as much sense, however, to say that Deep Throat could have been a butcher, Mark Twain could have been a baker, etc. I guess I think that for these sorts of names, something like a descriptive theory is needed.

I would appreciate if someone would expand on how primary and secondary intensions work in respect to definite descriptions and proper names.

Thanks- Rachael

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 12 12:36:43 1999
Date: Fri, 12 Mar 1999 13:33:09 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Angela J Burnette <aburnett@U.ARIZONA.EDU>
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

In our last meeting, Dave claimed that definite descriptions like 'the current president of the United States' would not refer to Bill Clinton in a poss. world where he was not the current prez even when that world was considered as counterfactual, is this because definite descriptions aren't rigid designators? Also, I'm confused (in these cases) as to what the primary and secondary intensions of such descriptions are like...because it seems as though, given that the referent of a description like 'the current president' doesn't pick out Bill Clinton in all poss. worlds *when considered as counterfactual*, that there...is no? secondary intension, or if there is, it's not serving the same...function? as the SI of a rigid designator (because it doesn't seem to fix the reference across worlds given that the actual world is the way it is). A little clarification here would be helpful...

Also, regarding the concept of a priori, or rather the definition that 'x is knowable a priori if x can be known prior to experience,' I don't see why problems like the one presented by Godel's theorem aren't more

troubling...if we have to idealize away from our own cognitive limitations to come up with an appropriate idea of 'in principle a priori knowability' such that we are conceiving of a being that is not subject to the sorts of mathematical counterexamples like Godel's theorem, then it seems as if our concept of a priori knowable breaks down...i.e., how are we to conceive of the distinction between what is knowable prior to experience and what is not for an idealized being like the one needed to avoid the Godel counterexample? It's not clear to me what such a being could or couldn't know without actually having to look...if a priori were defined strictly in terms of the analyticity of statements, then there seems to be less of a problem keeping the distinction between a priori and a posteriori straight when idealizing so far from our own case...but this doesn't seem to be the way Dave wants, centrally, to define the a priori...so is it the case that "knowable prior to experience" is more akin, or closer in kind to an idea like "analyticity" than I think, or what?

angela

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 14 15:29:45 1999
Date: Sun, 14 Mar 1999 15:29:28 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Primary and secondary intensions
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Angela, Rachael, and Erik L. raise a few lingering questions about the two-dimensional framework, primary and secondary intensions.

Angela writes about the secondary intensions of descriptions:

>In our last meeting, Dave claimed that definite descriptions like 'the
>current president of the United States' would not refer to Bill Clinton in
>a poss. world where he was not the current prez even when that world was
>considered as counterfactual, is this because definite descriptions aren't
>rigid designators?

Yes, this more or less comes down to the fact that descriptions aren't rigid designators. Let's take "the 42nd president of the United States" as our description to avoid difficulties due to indexicals and centering. And let's consider a world in which Paul Tsongas is the 42nd president of the United States -- and let's not consider it as actual (as an epistemic possibility), but as counterfactual (as a way the world might have been).

In that counterfactual world, what does "Bill Clinton" pick out? I.e., who is Bill Clinton in that counterfactual world. Certainly not Paul Tsongas! It's Clinton, who presumably is somewhere back in Arkansas in this world.

In that counterfactual world, what does "the 42nd president of the United States" pick out? I.e., who is the 42nd president of the United States in that world? Certainly not Clinton! It's Paul Tsongas, who is busy being president in this world (at least until he has a heart attack).

All this is borne out by the fact that we describe this as a counterfactual world in which Tsongas (not Clinton) is the 42nd president of the US. If "the 42nd president" picked out Clinton in this world considered as counterfactual, it would come out as a

counterfactual world in which Clinton is the 42nd president, which seems wrong.

This brings out Kripke's central distinctions between names and descriptions, which is that names are rigid designators and descriptions are not. That is, names pick out the same individual across all counterfactual worlds, but descriptions need not -- they pick out whatever individual satisfies the description, even in counterfactual worlds.

For example, "Clinton" picks out the same person in all worlds considered as counterfactual; i.e. the SI of "Clinton" picks out the same person in all worlds; i.e. "Clinton" is a rigid designator. Whereas "the 42nd president" picks out whoever is the 42nd president in a given world considered as counterfactual; i.e. the SI of "the 42nd president" picks out different individuals in different worlds; i.e. "the 42nd president" is not a rigid designator.

So the SIs of "Clinton" and "the 42nd president" are different. If they were the same, then "Clinton is the 42nd president" would be true in all counterfactual worlds, i.e. it would be necessary (in the Kripkean sense) that Clinton is the 42nd president, which seems wrong. The Tsongas world itself seems to bear witness to the fact that it is not necessary that Clinton is the 42nd president. We might say that SI of "Clinton" goes with the person, while the SI of the "42nd president" goes with the office.

There is a temptation sometimes to think that the secondary intension of a term is just *defined* as the function that picks out the actual referent (e.g., Clinton for "the 42nd president") in all worlds. (Even some well-known philosophers have interpreted the framework that way). Of course one could define an intension that works that way, but that isn't the way secondary intensions (as I've defined them) work. One simply takes a world, considers it as counterfactual, and sees what the term picks out there. Kripke's contribution was to make the nontrivial observation that names pick out the same referent in all worlds considered this way, whereas descriptions don't. That is, he made the nontrivial observation that names but not descriptions are rigid designators. If we defined secondary intensions in the alternative way suggested above, then it would be trivially true that names, descriptions, and everything else are rigid designators, so we would lose both the nontriviality of Kripke's observation.

Indeed, if we did things that way, we'd lose any way to distinguish between names and descriptions (such as "Hesperus" and "the evening star") at all. Their primary and secondary intensions would be the same, and both expressions would pick out the same things in all worlds. So we'd lose the ability to account for the fact that it is not necessary that Hesperus is the evening star, and so on. Or at least, we'd break the tie between secondary intensions and necessity, which would seem undesirable.

>Also, I'm confused (in these cases) as to what the
>primary and secondary intensions of such descriptions are
>like...because it seems as though, given that the referent of a
>description like 'the current president' doesn't pick out Bill Clinton in
>all poss. worlds **when considered as counterfactual**, that there...is no?
>secondary intension, or if there is, it's not serving the same...function?
>as the SI of a rigid designator (because it doesn't seem to fix the
>reference across worlds given that the actual world is the way it is). A
>little clarification here would be helpful...

The PI of "the 42nd president of the US" picks out the 42nd president in all worlds. The SI picks out the 42nd president in all worlds. The actual referent is Clinton.

The PI of "Bill Clinton" picks out very roughly "the guy I've heard of as 'Bill Clinton'" in all worlds. The SI picks out that very guy in all worlds. The actual referent is Bill Clinton.

This all reflects the fact that the SI of a name picks out the actual referent in all worlds, whereas the SI of a description doesn't. The description still has an SI, as one can still consider a counterfactual world and ask what the description picks out there. It's just that for a description, it needn't pick out the actual referent (as the term isn't a rigid designator). Basically, the SI of a name depends very heavily on the way the actual world turns out (it is very much a posteriori). But the SI of a description depends much less heavily on the way the actual world turns out (sometimes it may be a priori, or sometimes it may be somewhat a posteriori, e.g. due to the presence of a name within the description).

It's still the case that when we consider the Tsongas world as counterfactual, we're holding fixed that the actual world has turned out this way, with Clinton being such-and-such a guy, being the 42nd president, etc. Given that a posteriori information, this tells us that "Clinton" picks out that very guy across all worlds. But even this a posteriori information doesn't tell us that "the 42nd president" picks out Clinton across all worlds. Even knowing how the actual world turned out, the Tsongas world in question is described as a counterfactual world in which Tsongas is the 42nd president, not Clinton.

[All this is complicated a little by the fact that descriptions arguably have *a* reading on which they are rigid designators -- e.g. where one reads "the 42nd president" as something like "the actual 42nd president". On this reading, one could consider the Tsongas world as a world in which the 42nd president isn't president (i.e., the actual 42nd president, Clinton, isn't president), and on that reading, "the 42nd president" would pick out Clinton in the Tsongas world, not Tsongas (and would pick out Clinton in all worlds). But I think this would be a somewhat unusual reading of the description. At most, this would suggest that descriptions are ambiguous between a nonrigid and a rigid reading. The difference between the SIs of rigid terms and nonrigid terms would stay intact.]

Rachael writes:

>Like Angela, I am confused about what the secondary intension points to
>in the case of the forty-second president of the United States. To
>elaborate on a point that Brad made about Deep Throat in our last meeting,
>I can think of a few cases where rigid designators don't seem to work the
>way Kripke intends. For example, Mark Twain; it would seem, given in
>our actual world that Mark Twain is Samuel Clemmons, Mark Twain would be
>Samuel Clemmons in all counterfactual worlds (I guess Samuel Clemmons
>isn't really a secondary intension?). But I can easily imagine a world
>where Mark Twain was Joe Schmoe. Likewise, I can imagine a world where
>Marilyn Monroe was not Norma Jean and Aristotle was a woman. I guess this
>is possible because I am assuming that the primary intension of those
>names apply to Mark Twain, the author; Marilyn Monroe, the actress; and
>Aristotle, the philosopher.

Hmm, interesting. I think one can certainly imagine a world

considered as actual in which Marilyn Monroe wasn't Norma Jean, etc (just say we discovered that she was really born "Marilyn" and that the "Norma Jean" business was a myth, with the Norma Jean she pretended to be still living happily in Florida as a grandmother). But can we imagine a world considered as counterfactual in which Marilyn wasn't Norma Jean?

Take a counterfactual world where the real Norma Jean (who in the actual world became Marilyn) never went into show business, and that some other blonde actress (Jayne Mansfield, say) took the name "Marilyn Monroe", had some plastic surgery, ended up looking and sounding and living just like the actual Marilyn. Is this a counterfactual world in which Marilyn was Jayne Mansfield? Or is it a counterfactual world in which Marilyn never went into show business, and in which someone else looked and lived like her?

Kripke's intuition is the second, in which case the SI of "Marilyn Monroe" picks out Norma Jean in this world, not Jayne Mansfield. Do you have the first intuition instead? If so, then presumably the SI of "Marilyn Monroe" picks out Jayne Mansfield in this world, and picks out roughly whoever plays the relevant actress role in a given world.

My own intuitions tend to side with Kripke here. It does seem that we can consider counterfactual worlds in which Marilyn Monroe never went into acting, and lived a happy family life in Florida. If so, there are counterfactual worlds where the SI of "Marilyn" picks out Norma Jean, irrespective of whether she's an actress. And it is not obvious to me that we can easily consider a world where Jayne Mansfield really was Marilyn, as opposed to looking and living like her. If so, then the SI of "Marilyn" goes with the person, not the role.

But maybe you have different intuitions here. If you're right, there is at least one reading of some names in which they are nonrigid: "Marilyn Monroe" will pick out whoever plays the actress role in a counterfactual world, irrespective of whether they are the same person as the actual Marilyn. If this is right, then the distinction between names (at least these names) and descriptions would be broken down further.

Anyway, I'm not certain whether you're meaning to endorse these claims, or just the weaker primary intension claim (which applies only to worlds considered as actual). The weaker claim I think is very reasonable; the stronger claim would be unorthodox and many would disagree, but that's not to say that it's untenable. On this reading, I presume you'll have to disagree with Kripke, and argue that it is necessary that Hesperus is the evening star, in all counterfactual worlds, and so on?

>But Kripke is intent on showing that names are not descriptive. The name, >Marilyn Monroe, refers to, roughly, that woman that we call Marilyn >Monroe. I think that Kripke can make a good case in respect to Aristotle >(because Aristotle refers to the man and not to the Greek philosopher etc. >it makes sense to say that Aristotle could have been a blacksmith.) I >think that it does not make as much sense, however, to say that Deep >Throat could have been a butcher, Mark Twain could have been a baker, etc. >I guess I think that for these sorts of names, something like a >descriptive theory is needed.

>

>I would appreciate if someone would expand on how primary and secondary >intensions work in respect to definite descriptions and proper names.

Hmm, OK, so maybe you are endorsing the strong claim here. I think

Kripke would argue that Deep Throat could have been a butcher. (One can imagine Nixon saying "if only that damned Deep Throat had never gone into public affairs and had become a butcher instead".) And I think he would argue that "Deep Throat" can't pick out other individuals in other counterfactual worlds. (Of course it can pick out other individuals in worlds considered as actual, as the primary intension is pretty clearly descriptive.)

On your view, it looks like there will be two classes of names, some of which are rigid designators and some of which are not. The rigid ones pick out the same individuals across counterfactual worlds (they have a constant SI), whereas the nonrigid ones don't (they have a descriptive SI). I guess the nonrigid ones will be the ones that "look" descriptive in some sense, or which have strongly descriptive connotations -- Deep Throat, Jack the Ripper, Hesperus, etc, etc? Such names would presumably have very similar PI and SI, captured by the description in both cases. The rigid ones, by contrast (e.g. Rachael Parkinson, Joe Schmo, etc) will have some sort of subtle PI (maybe captured causally or metalinguistically) and will have an SI that picks out the same person in all worlds.

As I say, that would be an unorthodox but interesting position. (The orthodox position is that all names are rigid, so that their SI picks out the same individual in all worlds.) It might be tricky to draw the line between the two sorts of names. I guess an alternative would be to say that many or all names are ambiguous between a rigid and a nonrigid reading, so that "Deep Throat" etc can be read both ways (on one reading, Deep Throat could have been a butcher, on the other reading, not). Maybe one could even make the case for all names, so that e.g. "Rachael Parkinson" has an alternative descriptive reading on which Rachael had to be called 'Rachael' or some such. This would break down the barrier between names and descriptions still further, as we'd now have ambiguity between rigid and nonrigid readings in both cases.

All that applies to secondary intensions, of course. I think the primary intensions of names etc will be fairly unaffected by what we say here about rigidity.

Erik writes about causal primary intensions:

>I have a brief question regarding the admissability of causal primary
>intensions. This has come up at least twice; once in Tim's response to
>Josh's BIV scenario, and then again in our first class meeting. So, the
>problem is, I don't think primary intensions can ever be appropriately
>thought of as "the cause of my sensation that" or "that which brought
>about my experience of" or some such, because in any of these cases the
>primary intension can vary wildly, depending on what the cause happens to
>be. So some brain in a vat may have a PI of water that appeals to
>electrodes and stimulation or whatever, and on earth this of course won't
>make any sense. So "that which causes my x experience" makes a primary
>intension the same as whatever the cause happens to be, and not the
>essential notion or properties of the thing itself, apriori. So I think
>the primary intension has to be necessarily connected to the essential
>quality of the thing (at least as it is known apriori or conceptually).
>What we mean by the concept of water, then, is not what causes the water
>experience (which may have nothing to do with the essence of water itself).

Hmm, I'm not sure I've entirely grasped the problem here. It seems to me that the BIV case as you've described it is compatible with the PI of water being "that which causes my x experience". If that's the PI

for the BIV, then presumably "water" will pick out whatever causes x experiences for the BIV, which will presumably be some assortment of chemicals and electrodes, etc, which is just what you say.

Here there will be no difference between the PI of "water" in the BIV and in you and me. In both cases it might be "what causes my x experiences". Of course the **referent** will be different between the cases: for you and me, the PI will pick out H₂O, whereas for the BIV, it will pick out electrodes and such. But the PI and the referent are different things. One shouldn't strictly speaking say that the PI for the BIV involves electrodes, etc. It's important to remember that the PI is a function from worlds to extensions, and the referent is just the extension at the actual world. For the BIV, the PI might be "what causes my x experiences". One can say that the PI **picks out** electrodes and such in its world, or one can say that the **referent** involves electrodes, but these are different claims.

It may be that one wants to reject simple causal PIs for other reasons. E.g., one might reasonably argue that the term "water" for a BIV **doesn't** pick out the electrodes, even though they cause x experiences in the BIV. If that's right, then arguably the PI for "water" has a descriptive component over and above the pure causal component. But that's a subtle matter.

Let me know if I've missed your point here.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 12 14:58:13 1999
Date: Fri, 12 Mar 1999 15:54:59 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Rachael J Parkinson <rachaelp@U.ARIZONA.EDU>
Subject: Re: your mail
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Like Angela, I am confused about what the secondary intension points to in the case of the forty-second president of the United States. To elaborate on a point that Brad made about Deep Throat in our last meeting, I can think of a few cases where rigid designators don't seem to work the way Kripke intends. For example, Mark Twain; it would seem, given in our actual world that Mark Twain is Samuel Clemmons, Mark Twain would be Samuel Clemmons in all counterfactual worlds (I guess Samuel Clemmons isn't really a secondary intension?). But I can easily imagine a world where Mark Twain was Joe Schmo. Likewise, I can imagine a world where Marilyn Monroe was not Norma Jean and Aristotle was a woman. I guess this is possible because I am assuming that the primary intension of those names apply to Mark Twain, the author; Marilyn Monroe, the actress; and Aristotle, the philosopher.

But Kripke is intent on showing that names are not descriptive. The name, Marilyn Monroe, refers to, roughly, that woman that we call Marilyn Monroe. I think that Kripke can make a good case in respect to Aristotle (because Aristotle refers to the man and not to the Greek philosopher etc. it makes sense to say that Aristotle could have been a blacksmith.) I think that it does not make as much sense, however, to say that Deep Throat could have been a butcher, Mark Twain could have been a baker, etc. I guess I think that for these sorts of names, something like a descriptive theory is needed.

I would appreciate if someone would expand on how primary and secondary intensions work in respect to definite descriptions and proper names.

Thanks- Rachael

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 20 12:33:56 1999
Date: Tue, 20 Apr 1999 12:14:54 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: necessity of conditionals
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

I have an observation on indicative and subjunctive conditionals that I want to run by all of you. In every example that I have come up with if an indicative conditional has its truth value necessarily, then its corresponding subjunctive conditional also has its truth value necessarily. The truth values may differ, but it doesn't seem that the necessity of those truth values differs. Consider the following examples.

"If the CIA didn't arrange Kenedy's death then someone else did."
[Possibly true]

"If the CIA hadn't arranged Kenedy's death then someone else would have."
[Possibly false]

"If the stuff in the lakes and rivers is XYZ then water is H2O."
[Necessarily False]

"If the stuff in the lakes and rivers were XYZ then water would be H2O."

[Necessarily True]

"If Penguins don't live in Antarctica then my maps are misnaming the southern most continent in the world." [Possibly false]

"If Penguins hadn't lived in Antarctica then my maps would have misnamed the southern most continent in the world." [Possibly false]

Now this could be a fluke and someone may be able to generate a counter-example. But if there is a direct correspondence then surely it is because there is a connection between the necessity of indicative and subjunctive conditionals. And if there is a connection then one wonders whether we could do away with one or the other.

Just some thoughts to chew on.

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 20 22:08:50 1999

Date: Tue, 20 Apr 1999 22:05:53 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: necessity of conditionals

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: R

Josh writes:

>I have an observation on indicative and subjunctive conditionals that
>I want to run by all of you. In every example that I have come up
>with if an indicative conditional has its truth value necessarily,
>then its corresponding subjunctive conditional also has its truth value
>necessarily. The truth values may differ, but it doesn't seem that
>the necessity of those truth values differs. Consider the following examples.

We decided in the break today that the claim was as follows. Let S(IC) be the indicative conditional "If P is the case, then Q is the case", and S(SC) be the subjunctive conditional "If P were the case, Q would be the case." Then the claim is: S(IC) is indicatively necessary or indicatively impossible if and only if S(SC) is subjunctively necessary or subjunctively impossible.

I think I can come up with a few counterexamples, as follows. I've labelled them in Josh's way.

"If the tallest person in England committed those murders, the tallest person is Jack the Ripper".
[Necessarily true.]

"If the tallest person in England had committed those murders, the tallest person would have been Jack the Ripper".
[Probably false, but not necessarily.]

"If H2O is not in the oceans and lakes, water is XYZ".
[Possibly true, possibly false.]

"If H2O were not in the oceans and lakes, water would be XYZ".
[Necessarily false.]

It's interesting that many of the standard cases line up together in

Josh's way. It might be an interesting exercise to figure out what those examples have in common which the examples above do not (apart from the property at issue).

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 27 14:26:11 1999
Date: Tue, 27 Apr 1999 14:23:43 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Rachael J Parkinson <rachael@U.ARIZONA.EDU>
Subject: Indicative and subjective.
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

I just want to resolve a worry I expressed in seminar last week:

Indicative: If XYZ is in the oceans and lakes, then water is XYZ.

Subjunctive: If XYZ were the liquid in the oceans and the lakes, water wouldn't be.

I argued that it made sense to say "If XYZ were the liquid in the oceans and lakes, water would be XYZ." That is to say, in considering this world as actual, we could discover that we were mistaken in identifying water with H₂O, that it is actually XYZ. The sentence looks like a subjunctive one.

The mistake here is that the sentence is actually indicative, roughly equivalent to "If we were to find out that XYZ is in the oceans and lakes, then water is XYZ."

Still, I have a confusion which I'm sure someone will be able to easily dispel. Chalmers suggested that we can match the subjunctive up with secondary intensions and the indicative with primary intensions. It looks like the indicative sentence "If XYZ is in the oceans and lakes, then water is XYZ" applies to the secondary intension of water.

What do you think?

-Rachael

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 27 17:14:44 1999
Date: Tue, 27 Apr 1999 17:12:53 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: Re: Indicative and subjective.
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

I'm not sure if this will be helpful, but here goes...

>

> Indicative: If XYZ is in the oceans and lakes, then water is XYZ.

> Subjunctive: If XYZ were the liquid in the oceans and the lakes, water wouldn't be.

> I argued that it made sense to say "If XYZ were the liquid in the oceans and lakes, water would be XYZ." That is to say, in considering this world as actual, we could discover that we were mistaken in identifying water with H₂O, that it is actually XYZ. The sentence looks like a subjunctive one.

I think you mean to say that this is an indicative conditional-- considering the centered world in which XYZ is the dominant watery stuff, water is XYZ.

> The mistake here is that the sentence is actually indicative,
roughly
> equivalent to "If we were to find out that XYZ is in the oceans and lakes,
> then water is XYZ."
> Still, I have a confusion which I'm sure someone will be able to easily
> dispel. Chalmers suggested that we can match the subjunctive up
> with secondary intensions and the indicative with primary intensions. It
> looks like the indicative sentence "If XYZ is in the oceans and lakes,
> then water is XYZ" applies to the secondary intension of water.
> What do you think?
> -Rachael
>

It seems to me that this last sentence, "If XYZ is in the oceans and lakes, then water is XYZ", is making a statement about the primary intension of water. The way we pick out water according to the primary intension is that it is whatever plays the role of being the dominant watery stuff in a centered world. If we consider the secondary intension, however, water is H₂O. Accordingly, a counterfactual world with XYZ in the rivers and lakes is actually a world without water.

It seems that you can construct these conditionals using terms that make it sound that an indicative cond. is a subjunctive and the reverse. It seems, perhaps, that one needs to simply think about indicative conditionals as those that are saying something like, 'assuming that scu and such actually is the case, x follows'. And, similarly, subjunctives are those conditionals that ask you to consider what would follow if something that is not the case, was.

sorry if this is not helpful.
Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Wed Apr 28 10:42:08 1999
x-sender: agillies@pop.u.arizona.edu
Date: Wed, 28 Apr 1999 10:56:21 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: truth conditions of indicatives
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

A lot of people talking about indicatives use the Ramsey Test to get assertability conditions for indicatives. But to ground truth₁ and truth₂ (and therewith Nec₁ and Nec₂) we need bona fide truth conditions. The worry is that the Ramsey Test brings into play other stuff in our belief corpus, and so might not be capable of fixing truth conditions. Here's one way that the Ramsey Test for indicatives can get going on genuine truth conditions.

Say that a world w₂ is epistemically accessible to an agent in w₁ just in case for all the agent knows, w₂ might be the actual world. Accessibility of epistemic alternatives is an equivalence relation. An agent knows P iff P is true in all the epistemically accessible worlds to the agent. Now let's beef up the notion by saying that in this sense of 'knows', an agent knows all the consequences of what she knows. As a limiting case, some agents can know everything there is to know at their worlds. Let K be such an agent at w. Further, take worlds to be sets of propositions. Then:

The indicative "P --> Q" is true at w iff: for any w' such that w' is

epistemically accessible from w to K, the minimal revision of w' to include P includes Q.

I think this gets out of the worry.

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 29 14:08:49 1999
Date: Thu, 29 Apr 1999 13:54:43 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Indicative and subjunctive
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Rachael wrote:

>I just want to resolve a worry I expressed in seminar last week:
>
>Indicative: If XYZ is in the oceans and lakes, then water is XYZ.
>Subjunctive: If XYZ were the liquid in the oceans and the lakes, water
>wouldn't be.
>
>I argued that it made sense to say "If XYZ were the liquid in the oceans
>and lakes, water would be XYZ." That is to say, in considering this world
>as actual, we could discover that we were mistaken in identifying water
>with H2O, that it is actually XYZ. The sentence looks like a subjunctive
>one.
>
>The mistake here is that the sentence is actually indicative, roughly
>equivalent to "If we were to find out that XYZ is in the oceans and lakes,
>then water is XYZ."

Hmm, I'm not sure that that sentence is grammatical! It looks like a mixture of subjunctive and indicative to me.

One could try making it more indicative all the way by going to:

"If we find out that XYZ is in the oceans and lakes, then water is XYZ".

Alternatively, one express something like this with a "metalinguistic subjunctive":

"If XYZ were the liquid in the oceans and lakes, we would say that 'water' refers to XYZ."

Arguably, metalinguistic subjunctives like this, which *mention* the term 'water' rather than *using* it, come close to tracking the primary intension of 'water' rather than the secondary intension (though there are subtleties here).

>Still, I have a confusion which I'm sure someone will be able to easily
>dispel. Chalmers suggested that we can match the subjunctive up
>with secondary intensions and the indicative with primary intensions. It
>looks like the indicative sentence "If XYZ is in the oceans and lakes,
>then water is XYZ" applies to the secondary intension of water.

I think I agree with Anthony's analysis of this. It seems to me that your (Rachael's) sentence tracks the primary intension of 'water', not

the secondary intension. Taking the secondary intension, 'water is XYZ' will be false in a scenario where XYZ is in the oceans and lakes. But taking the primary intension, it will be true there. Insofar as we judge the indicative conditional above to be true (or assertible), it seems to mirror the primary intension.

Anthony wrote:

>It seems that you can construct these conditionals using terms that make
>it sound that an indicative cond. is a subjunctive and the reverse. It
>seems, perhaps, that one needs to simply think about indicative
>conditionals as those that are saying something like, 'assuming that scu
>and such actually is the case, x follows'. And, similarly, subjunctives
>are those conditionals that ask you to consider what would follow if
>something that is not the case, was.

Right, an alternative is to think about the conceptual categories of "epistemic conditionals" and "counterfactual conditionals", or some such, rather than directly in terms of the grammatical categories. Epistemic conditionals are those that consider the antecedent as actual, and counterfactual conditionals are those that consider the antecedent as counterfactual (more or less as you suggest above). It then turns out that most of the time, grammatically indicative conditionals are epistemic conditionals, and grammatically subjunctive conditionals are counterfactual conditionals. But maybe there can be cases or readings where this is not so. In any case, it's the conceptual rather than the grammatical distinction that runs deepest.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 29 15:24:04 1999
Date: Thu, 29 Apr 1999 14:06:29 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: truth conditions of indicatives
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Thony wrote:

>A lot of people talking about indicatives use the Ramsey Test to get
>assertability conditions for indicatives. But to ground truth_1 and
>truth_2 (and therewith Nec_1 and Nec_2) we need bona fide truth
>conditions. The worry is that the Ramsey Test brings into play other
>stuff in our belief corpus, and so might not be capable of fixing truth
>conditions. Here's one way that the Ramsey Test for indicatives can get
>going on genuine truth conditions.

Well, my hope is that we needn't take a stand on whether typical indicative conditionals have truth-conditions or just assertibility-conditions. As we discussed in class last week, the belief-relativity seems to apply only to the way a partial antecedent gets fleshed out into a whole world, rather than to the way that world gets evaluated. And truth_1 just needs the world-evaluation. So even if indicative conditionals with partial antecedents have belief-relative truth- or assertibility-conditions, an objective sort of semantic evaluability across worlds seems to lie in the background.

Still, it would be nice to have some sort of non-relative truth-conditions for indicatives with partial antecedents that comes close to capturing some of the intuitive assertibility-conditions.

So let's see.

> Say that a world w_2 is epistemically accessible to an agent in w_1 just in
> case for all the agent knows, w_2 might be the actual world.
> Accessibility of epistemic alternatives is an equivalence relation. An
> agent knows P iff P is true in all the epistemically accessible worlds to
> the agent. Now let's beef up the notion by saying that in this sense of
> 'knows', an agent knows all the consequences of what she knows. As a
> limiting case, some agents can know everything there is to know at their
> worlds. Let K be such an agent at w . Further, take worlds to be sets of
> propositions. Then:
>
> The indicative " $P \rightarrow Q$ " is true at w iff: for any w' such that w' is
> epistemically accessible from w to K , the minimal revision of w' to
> include P includes Q .

Interesting proposal! So basically, the truth-conditions of an indicative will correspond to the assertibility-conditions for an omniscient being. My main worry is that it is not obvious that there is an objective notion of "minimal revision" for an omniscient being. If a being who knows everything (or thinks they do) finds out that their belief that not- P is false, how will they revise? Presumably there will be lots of different ways to revise. And presumably any such way will require giving up on some other beliefs, and preserving others. The decision between these seems to turn on which linked beliefs are "strongest" or closest to the "core" for such a being.

It seems to me that two omniscient being might well give two quite different judgments about the indicative above, depending on just which of their beliefs that view as most amenable to revision. If so, the problem of relativity arises once again.

Of course this problem won't arise for indicatives with whole worlds in the antecedents. In this case, there won't be any "slack" in the belief revision process, and the outcome should be determined.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 29 15:26:35 1999
Date: Thu, 29 Apr 1999 14:40:50 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: Re: truth conditions of indicatives
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Thony and Dave wrote:

> > Say that a world w_2 is epistemically accessible to an agent in w_1 just in
> > case for all the agent knows, w_2 might be the actual world.
> > Accessibility of epistemic alternatives is an equivalence relation. An
> > agent knows P iff P is true in all the epistemically accessible worlds to
> > the agent. Now let's beef up the notion by saying that in this sense of
> > 'knows', an agent knows all the consequences of what she knows. As a
> > limiting case, some agents can know everything there is to know at their
> > worlds. Let K be such an agent at w . Further, take worlds to be sets of
> > propositions. Then:
> >
> > The indicative " $P \rightarrow Q$ " is true at w iff: for any w' such that w' is
> > epistemically accessible from w to K , the minimal revision of w' to
> > include P includes Q .

>
> Interesting proposal! So basically, the truth-conditions of an
> indicative will correspond to the assertibility-conditions for an
> omniscient being. My main worry is that it is not obvious that there
> is an objective notion of "minimal revision" for an omniscient being.
> If a being who knows everything (or thinks they do) finds out that
> their belief that not-P is false, how will they revise? Presumably
> there will be lots of different ways to revise. And presumably any
> such way will require giving up on some other beliefs, and preserving
> others. The decision between these seems to turn on which linked
> beliefs are "strongest" or closest to the "core" for such a being.
>
> It seems to me that two omniscient being might well give two quite
> different judgments about the indicative above, depending on just
> which of their beliefs that view as most amenable to revision. If so,
> the problem of relativity arises once again.

I took Thony to be saying that minimal revisions are not belief revisions, but world revisions. The idea is to get the nearest possible P world to w. So for Thony's case you first pick out the equiv. class of epistemically accessible worlds. Then look at all w' such that w' is an epistemically accessible world. P-->Q is true at w if for each w', the nearest P world to w' is also a Q world. It doesn't matter how K will update or revise his beliefs because "nearest P world" is a relation between worlds and is independent of agents.

Now you might argue that the notion of a "nearest P world" isn't clear. In fact I think it isn't at all clear. But pretty much everyone that talks about possible worlds wants to say that there is some such relation. So long as you hold that possible worlds are not themselves dependent on agents then any such relation is going to be objective.

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 29 15:47:21 1999
Date: Thu, 29 Apr 1999 15:45:57 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: truth conditions of indicatives
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

OK, so Josh suggests appealing to a non-epistemic notion of "nearness" between worlds to evaluate indicatives. I take it this is supposed to be the same nearness relation that is present in evaluating partial subjunctive conditionals, e.g. on the Stalnaker/Lewis analysis (as Josh points to the fact that everyone accepts this relation already).

If this worked, it would give a nice link between the truth-values of indicatives and subjunctives. The subjunctive "If P, then Q" will be true iff the nearest P-worlds are Q-worlds, where P-worlds etc are worlds satisfying P's secondary intension. The indicative "If P, then Q" will be true iff the nearest P-worlds are Q-worlds, where P-worlds etc are worlds satisfying P's primary intension.

I'm not sure that this is just what Thony intended, since the idea of an omniscient being now seems to drop out of the analysis. But in any case, I worry that the same nearness relation can't plausibly do the job in both cases. Take the conditional with P = "Butterfly B flapped

its wings like so last week", and Q = "It rained this week". It may well be that the subjunctive "If P then Q " is true -- if the butterfly *had* flapped its wings, it would have rained this week. (I'm assuming it didn't actually rain this week.)

But it doesn't seem right to say that the indicative "If P , then Q " is true. If we were to discover that the butterfly flapped its wings, we wouldn't suddenly decide that it rained. And even an omniscient being probably wouldn't reason in that way, it seems. After all, that being will have a whole lot of beliefs about the way things are this week that would have to be overturned just to accommodate the tiny change in beliefs about the butterfly. It seems more plausible to say that the being would make a few local adjustment in its beliefs about the way things are in the vicinity of the butterfly, in order that it can change those beliefs without having huge effects elsewhere (maybe adjusting beliefs about a few molecules in the vicinity to compensate).

The intuitive moral, I think, is that indicative conditionals require something like "epistemic nearness", which is somewhat different from the sort of "metaphysical nearness" required for subjunctives. Of course one could always just stipulate the latter, but then one is moving a fair distance away from the intuitive correctness conditions for indicative conditionals.

--Dave.

P.S. Re Kripke/Wittgenstein on the social community, I tend to think that this is hopeless as a "solution" to the puzzle, as any in-principle argument that demonstrates indeterminacy in an individual will also demonstrate indeterminacy in a community. After all, in relevant respects, the community can be seen as a big individual, or an individual can be seen as a little community.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 19 06:21:34 1999
Date: Fri, 19 Feb 1999 06:20:07 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Next week
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

My visa still isn't through, and although there are indications that it shouldn't be too long now, I won't be there by next Tuesday. This means I have to rearrange things a little. Week 5 was originally supposed to be on the "tyranny of the subjunctive", on relationships between the 2-D framework and indicative and subjunctive conditionals, and on implications for the analysis of necessity. But that isn't really written up (except very briefly in the Princeton outline); I'd been planning to mostly talk about it in class. So as things stand, we will skip straight to the following topic (the original week 6) instead.

This topic involves applying the 2-D framework to epistemic and modal arguments against materialism. The readings are

TCM, Chapter 4 -- mostly pp. 131-149.
N&N, pp. 144-55.
Mind and Modality, Lecture 1 (esp. sections 5-7).

The Mind and Modality notes (on the web) have my currently preferred formulation of the 2-D argument against materialism. It's a bit crisper and tighter than the book version, though it comes to much the same thing in the end. The book version has more discursive detail, so you'll probably want to look at the two of them together (the real core of the book argument is pp. 131-36, but pp. 147-49 also has relevant material). It might also be interesting to look at Kripke's anti-materialist argument in N&N, which is similar in spirit, and compare and contrast. (N.B. The Princeton notes have some contrastive analysis that goes a bit beyond what is in the book.)

Anyway, consider these the readings for next week's discussion. I'll look forward to seeing your thoughts on them early next week.

As promised, there's also an assignment. I'd like you to take two terms that Kripke discusses, and translate his discussion into the 2-D framework. One is "yard", as discussed on p. 76; the other is "cat", as discussed on p. 122 and pp. 125-6. I'd like you to take every significant part of Kripke's discussion here and translate it into the 2-D framework. Characterize roughly what the PI and the SI of these terms look like, at least given Kripke's intuitions. In places where Kripke considers a hypothetical scenario, you should roughly specify the world, say whether it is being considered as actual or as counterfactual, note what the referent of the relevant term or statement is in that world (considered the relevant way, and according to Kripke's stated intuitions), and say what the upshot is the for relevant PI or SI. Try to translate Kripke's more general points in these passages into the framework too, if you can. If you're so inclined, you can say whether you agree or disagree with Kripke's specific and general analyses and why, though that isn't compulsory.

This shouldn't be much work (the passages are pretty short). This will be due by next Tuesday at noon, Arizona time. Don't be late. E-mail it directly to me (not to the mailing list). It should be your own work.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 24 03:11:57 1999
X-Accept-Language: en
Date: Wed, 24 Feb 1999 04:04:49 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Brad Thompson <bradt@U.ARIZONA.EDU>
Subject: Kripke and the primary intension of "pain"
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

On pp. 150-5 of N & N, Kripke argues that the identity theorist who claims that "pain = C-fiber stimulation" is necessary a posteriori cannot explain away the apparent contingency of the identity statement in the same way as with "water is H2O" or "heat is molecular motion". In the latter cases, Kripke claims that the apparent contingency is due to the falsity of a qualitatively analogous statement in a qualitatively identical epistemic situation. So, for example, strictly speaking it is false that heat might not have been molecular motion. But what is true is that someone might have the sensation of heat (that is, be in a qualitatively identical epistemic situation as the one I am in when I sense heat/molecular motion) even though the phenomenon being sensed is not molecular motion.

"Pain = C-fiber stimulation" can't be handled in the same way because, Kripke says, being in a qualitatively identical epistemic situation as the one in which I am having pain *is* to have a pain. In the end I might find this satisfying, but it does seem to me worthy of debate/discussion whether or not it is correct to here identify pain with an "epistemic situation". Pains are not conceptual, and there is lots that needs to be said about how they could be properly characterized as constituting a person's "epistemic situation". There is a sense in which it seems that a person (let's call her Jones) who lacked the capacity to feel heat sensations but who had a thermometer built into parts of her body which sent temperature information to her brain could be in the same epistemic situation as a "normal person". That is, she could be in equivalent information states.

In the Mind and Modality notes, Chalmers rejects the "epistemic situation" bit in the above for different reasons. But using the notion of a primary intension, I suppose that one would say that Jones does not have a concept of heat assuming that the actual sensations of heat are part of the PI of "heat". Does this seem right? It probably does, but let's suppose that Jones' judgments about the presence of heat is identical to mine in every way (she makes exactly the same discriminations, for example). I can imagine someone (perhaps a Dennett) arguing that Jones does have our concept of heat, and that the PI of heat involves something that Jones and I share (but which I find difficult to characterize).

Perhaps something similar could be done with "pain". Might this not defuse the argument against materialism? Of course, it depends on a highly contentious analysis of the PI of phenomenal concepts. At any rate, how we are to characterize the PI's of phenomenal concepts seems to be an interesting issue and one that is extremely important in assessing the anti-materialist argument.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 25 00:56:29 1999
Date: Thu, 25 Feb 1999 00:56:17 -0800
Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: Kripke and the primary intension of "pain"

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Brad raises a few interesting questions re "pain" and "heat". First, he wonders whether it's really right to say our epistemic situation re pain really includes pain. Second, he asks whether someone really needs to have heat sensations to have the concept of "heat" and the associated PI. Third, if the answer to this question is no, he wonders whether the point might be extended to "pain", so someone could have the concept and the PI without having sensations of pain.

(1) Re the first question, of course I think that epistemic situations are inessential here, and that nothing really turns on them in the argument. But I do share Kripke's intuitions that there's a reasonable sense in which when I'm feeling pain, pain is an essential part of my "epistemic situation". Pain is unlike chairs in that way, for example; when I look at a chair, it's arguable that I could be in the same "epistemic situation" in some sense without there being a chair there. This is just what skeptical arguments trade on -- the idea that my "direct evidence" or some such could be as it is for me right now, without various objects in the external world existing.

We might think of "epistemic situation" as being implicitly defined as the situation that is preserved in skeptical scenarios. I can consider a skeptical scenario where I'm in the same epistemic situation but there's no chair, but it's less clear that there's a skeptical scenario where I'm in the same epistemic situation but there's no pain, or no sensation of blue. Arguably, changing those things changes my "direct evidence", and changes the way things are for me now. Of course, this is an intuition and is somewhat loose, but it does seem to get at something central in our thinking about skeptical scenarios and what is held constant.

Personally, I think this gets at some deep differences between the epistemology of conscious experience (in the first-person case) and that of the external world. Somehow our epistemic relation to experience seems more direct and intimate than our epistemic relation to external objects. But that's a long story in its own right (there's a little about this in chapter 5 of the book and in the third M&M lecture).

Of course such views aren't uncontroversial. Some might argue that we really stand in the same epistemic situation vis-a-vis experience as vis-a-vis chairs. And some might argue that if experiences aren't conceptual, it's not really right to speak of them as being "epistemic" -- maybe our epistemic situation should be restricted to our beliefs, or some such. This might be particularly attractive to someone who thinks that the only thing that can justify a belief is another belief (a coherentist as opposed to a foundationalist, for example). Personally, I think there's a very strong intuition that experiences serve as basic evidence and can serve to justify beliefs even if they aren't conceptual, but there is certainly a lot that needs to be worked out there.

Brad wonders whether "Jones" who lacks heat sensations and instead has thermometer readings sent straight to her brain could be said to be in the same epistemic situation. Interesting. I guess I'd say that if the information is all unconscious and unexperienced for Jones, then Jones can consider skeptical scenarios in which the information is

different, etc., and can't rule out the existence of those situations in some intuitive sense, e.g. can't rule them out by "direct evidence". So there is a temptation to say that this info isn't part of the epistemic situation, at least not in the direct way that heat sensations are. Something to do with the skeptic-proofness of one but not the other. Still, there's some sense in which one might want to say that the temperature readings are "evidence" in some pretty direct sense. Maybe to a reliabilist, they'll be as direct evidence as anything (I don't know just what a reliabilist will say here, but maybe others can help me out). So anyway, there are some deep epistemological issues lurking around here.

(2) Do we need to have heat sensations to have the concept and the PI of "heat". Well, on the Kripkean view, the PI of "heat" is more or less "the cause of heat sensations", so the question of whether one can have this concept without heat sensations reduces to the question of whether one can have the concept (PI) of "heat sensations" without heat sensations. That one in effect was question (3).

In suggesting that Jones might have the concept of "heat", I can see three things one might be saying. (a) Jones has the PI "the cause of heat sensations" without heat sensations; (b) The PI of heat is something other than "the cause of heat sensations", and Jones has that PI; or (c) The PI of *our* concept of heat is "the cause of heat sensations", but Jones has a different PI that still qualifies as a concept of "heat".

I'm not sure which of these Brad was pointing to, but all are interesting. Option (a) (which reduces to the pain-like case from question (3)) doesn't seem too plausible here, since heat doesn't cause any heat sensations in Jones, so this PI wouldn't pick out anything. But (b) and (c) are left. I think it's plausible that the PI of Jones' "heat" is something like "the cause of heat judgments", or some such (I'm assuming she is led to the judgments in a blindsight-like way). The real question then is whether that counts as a concept of heat, and as the same as our concept.

One possibility is that the PI of our own concept is "the cause of heat judgments", not "the cause of heat sensations". That's interesting and not obviously false. I guess it comes down to: what do we say about a (broadly) epistemically possible situation in which we have no heat sensations but still have 'heat' judgments (e.g., the broad epistemic possibility that we are like Jones). If that turns out to be actual, will we say that heat is the stuff that causes the judgments, despite the lack of sensations? Maybe we would. If so, the PI of *our* concept is arguably judgment rather than sensation oriented, and Jones can have that PI. This could be the "Dennett"-like option that Brad gestures towards. It's tricky, though. (I mention this sort of possibility in footnote 41 on p. 368 of the book.)

Alternatively, we might say Jones' PI is different from ours, but similar enough that it still counts as a concept of heat. Maybe so, though that would be a largely terminological issue. Personally, I think the previous possibility is more interesting.

(3) Can something similar be done with "pain"? Hmm. First, I note that the Jones case doesn't lend direct support, since we saw that that case can't rest on her having the PI "heat sensations" without heat sensations. But maybe something like this could happen anyway.

Possibilities: (1) Someone could have our PI of "pain" without having

pain sensations. (2) Someone could have a different PI of "pain" without pain sensations, but which would still be a concept of pain.

I have to go out now, so I'll get back to this later. All thoughts are welcome.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 25 22:10:12 1999
Date: Thu, 25 Feb 1999 22:09:46 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Kripke and the primary intension of "pain"
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

On Brad's third point re "pain" etc (continuing from last time). We were talking about the possibility that the PI of "heat" might not essentially involve heat sensations (e.g. maybe just "heat" judgments), and then wondering whether that might be extended to the concept of pain.

[(3) Can something similar be done with "pain"? Hmm. First, I note that the Jones case doesn't lend direct support, since we saw that that case can't rest on her having the PI "heat sensations" without heat sensations. But maybe something analogous could happen anyway?

Possibilities: (1) Someone could have our PI of "pain" without having pain sensations. (2) Someone could have a different PI of "pain" without pain sensations, but which would still be a concept of pain.]

I take it the interesting possibility is (1), as we are interested in the PI of our concept. Could Jones have e.g. our PI of "pain", despite her lack of sensations, in the way that she arguably has the PI of "heat"?

One move might be to suggest that the PI of "pain" should be analyzed as something like "the cause of 'pain' judgments", as in the heat case. If so, then Jones could truly say "I'm in pain", despite her lack of true sensations. Is our PI like this? This depends on what we'd say about the broad epistemic possibility that we are like Jones: what if that turned out to be actual? Maybe there's at least some sense in which we'd say that we still have pain, because we'll use "pain" for what causes the "pain"-judgments. That's not entirely clear to me, but maybe.

In any case it seems that one can just re-raise the issue with the concept of "pain qualia". I think it is relatively clear that the (broad) epistemic possibility that we are like Jones is an instance of the (broad) epistemic possibility that we don't have pain qualia. That is, the concept of "pain qualia" will *not* apply in the Jones centered world. (In a world of people like Jones, Dennett-ish people saying "pain qualia don't exist" would be right.) So the PI of "pain qualia" applies *essentially* to qualia-like things.

This in effect reflects Kripke's observation re "pain". Perhaps one could argue about it re "pain", but then one can just reraise it by explicitly building in "pain feeling" or "pain qualia" or some such into the concept -- i.e. by really specifying that it is a truly phenomenal concept. Once one has done that it seems we have a qualia whose PI applies essentially to pain qualia or something in the

vinicity. After all, we need *some* concept to distinguish our own actual situation from the centered world (considered as actual) in which we are something like Jones. (Unless one denies the distinction, in which case one is presumably denying the epistemic intuitions re qualia, zombies, etc.) And "pain qualia" seems to be just the concept. So beyond a certain point, the trick of pushing things out into the world seems to stop.

At least that's how it seems to me, but I may be wrong about this, so I'd be interested to hear more thoughts. This ties closely to the sort of issue about the content of phenomenal concepts in "The Content and Epistemology of Phenomenal Belief". There I argue that our most crucial phenomenal concepts have the same PI and SI, each picking out the appropriate sort of quality in all worlds. So the structure of phenomenal concepts is indeed special in a way. At least that's what falls out of accepting the epistemic intuitions, re the conceivability of inversions, zombies, etc. If those intuitions are rejected, it may be a different story.

Brad says that this issue re phenomenal concepts impacts the anti-materialist argument in important ways. I think he's right about that, but I'm interested to know how, exactly. E.g., how would it affect the argument I put forward, essentially from the two premises, (1) the epistemic gap and (2) the 2-D analysis of a posteriori necessity? It's not obvious how it affects premise (2). Are you suggesting that it affects premise (1), maybe? If so, I'm interested to hear more.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 25 22:46:43 1999

Date: Thu, 25 Feb 1999 23:45:49 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Josh Cowley <josh@MATH.ARIZONA.EDU>

Subject: cognitive deficits

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Keeping in line with my subject title the following was written while suffering from some cognitive deficits. I'm hoping the reaction of others will help me figure out what I'm worried about here.

I have a question about the section on cognitive limitations (pp. 138-140) Dave says that there is a weak sense in which we can conceive of a world in which the continuum hypothesis is true. But, since we don't know if it is true, it isn't the strong sense of conceivability that we get when we consider a physically identical world without consciousness. I'm wondering why this sense of conceivability is "weak."

I also want to push a worry that Rachael and I were expressing during discussion the other day. Kripke is of the opinion that there is only one actual world, this one. PIs pick out referents at centered possible worlds considered as actual. I take it, that to do this you need to be able to consider any possible world as actual. (First question: Is my last sentence right?) If this is right then consider a world in which I don't exist. (Index I to yourself). I can easily imagine a possible world in which I don't exist (a very unfortunate world). But I'm having trouble considering that world as actual. Now there are some ways of defining actual worlds such that I can consider a world in which I don't exist as actual. It seems like these

definitions do one of two things. Either they make me think of different possible worlds as counter-factual or they make me think of lists of propositions true at this world. At the moment I'm unable to articulate what I mean by that second choice. I'll try and fill it out tomorrow.

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Thu Feb 25 23:30:57 1999
Date: Thu, 25 Feb 1999 23:30:45 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: cognitive deficits
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

A quick reply to Josh.

>I have a question about the section on cognitive limitations
>(pp. 138-140) Dave says that there is a weak sense in which we can
>conceive of a world in which the continuum hypothesis is true. But,
>since we don't know if it is true, it isn't the strong sense of
>conceivability that we get when we consider a physically identical
>world without consciousness. I'm wondering why this sense of
>conceivability is "weak."

This comes down to the distinction between positive and negative conceivability that we will be discussing. Intuitively, we can't rule out the possibility that CH is true, but nevertheless we can't clearly and distinctly conceive of a scenario in which it is true. Negative conceivability goes with not finding a contradiction; positive conceivability goes with clear and distinct conception of a world, in some intuitive sense. Intuitively, we can flesh out a whole world in the "conceptual imagination" in some sense in which there are zombies, whereas we can't do anything like that fleshing out a world for CH (or not CH). It's more the purely formal inability to rule out the hypothesis.

If you like, think of positive conceivability as being able to conceive of a world verifying P, and negative conceivability as being unable to rule out P a priori. With zombies, we conceive of two distinct worlds, one with consciousness and one with zombies. But we don't have an intuition of conceiving of two distinct worlds, one with CH and the other with not-CH, in the mathematical case. So only the zombie case goes with a distinct positive conceivability intuition.

There's much more to say about this, but this is a start.

>I also want to push a worry that Rachael and I were expressing during
>discussion the other day. Kripke is of the opinion that there is only
>one actual world, this one. PIs pick out referents at centered
>possible worlds considered as actual. I take it, that to do this you
>need to be able to consider any possible world as actual. (First
>question: Is my last sentence right?)

That's more or less right. It may be that some worlds don't yield determinate extensions or truth-values, though (especially particularly distant worlds, or worlds where certain background assumptions aren't satisfied -- see "The Components of Content").

>If this is right then consider

>a world in which I don't exist. (Index I to yourself). I can easily
>imagine a possible world in which I don't exist (a very unfortunate
>world). But I'm having trouble considering that world as actual.

By "consider a world in which I don't exist", there are two things you can mean. Basically, one satisfying the PI of "I don't exist" or the SI. The first will be a world such that when it's considered as counterfactual, the verdict is that I don't exist. I.e. an instance of the subjunctive possibility that I don't exist. Essentially, that's a world in which this very person, David Chalmers, doesn't exist. The second is a world such that when it is considered as actual, the verdict is that I don't exist. I.e. an instance of the epistemic possibility that I don't exist. That's a very weird world, but in essence it is a centered world without anyone at the center of that world.

I'm not sure which of these you intend here, but my guess is that it's the first. If so, that's a slightly odd thing to do, as you're mixing "considering as counterfactual" (in specifying the world) with "considering as actual" (in evaluating the world). So there is the potential for getting mixed up. Still, if you're very careful, I guess you can do that. I can take worlds in which David Chalmers doesn't exist and consider them as actual. E.g., I can take a world centered on Michael Jordan and without David Chalmers. Considering that world as actual, it's more or less an instance of the epistemic possibility that I am Michael Jordan and that David Chalmers doesn't exist (or perhaps of the epistemic possibility that I am Michael Jordan and that the guy with such-and-such DNA structure etc doesn't exist). It's tricky, but one can do it.

What makes it so tricky of course is the way that one combines considering as counterfactual and as actual here. In a way, it's best to specify the world one is considering using "considering as actual" terminology all along. For example, one can just talk of the epistemic possibility that P, for a whole lot of P, in specifying a world. E.g. the very odd epistemic possibility that I don't exist. (It's disputable whether this is really a broad epistemic possibility, but I tend to think that it is, as it isn't a priori that I exist, it is a posteriori. That's a tricky matter, though.) Then I can say, if that world (the world with no-one at the center) is actual, I don't exist. So one can evaluate at least some terms and statements there.

>Now

>there are some ways of defining actual worlds such that I can consider
>a world in which I don't exist as actual. It seems like these
>definitions do one of two things. Either they make me think of
>different possible worlds as counter-factual or they make me think of
>lists of propositions true at this world. At the moment I'm unable to
>articulate what I mean by that second choice. I'll try and fill it
>out tomorrow.

Interesting, maybe these correspond in a way to the above. Your first choice seems to be the mixed strategy mentioned above (with both considering as counterfactual and as actual). I'm not sure about the second choice, but maybe it is like the second option above. This is all a tricky business so I'm interested to hear more.

--Dave.

P.S. I still need to see the minutes from this week's meeting. And lots of people still owe their comments re the week 5 readings and issues. We are falling behind a bit, so I hope people can do this

very soon.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 26 10:13:13 1999
Date: Fri, 26 Feb 1999 11:11:24 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: minutes
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Here are the minutes from this weeks meeting, compiled by Rachael and myself:

First, we talked about different positions regarding discussion of possible worlds. Kripke seems to suggest that we cannot make sense of considering other worlds as the **actual** world. Once we have identified, for instance, that water is H₂O, it does not make sense to consider worlds in which water is something else. This is distinct from considering the epistemic possibility of discovering that water actually is not H₂O in this world. We then debated the plausibility of considering other possible worlds as actual-- in short, whether PIs are, in fact, of primary importance to the meaning of a concept.

Then we discussed epiphenomenalism. Some members of the class thought that it is possible to accept the possibility of inverted spectra while resisting the notion that phenomenal experiences are causally inefficacious. Tim suggested that this is not the case, and that inverted spectra leads one down the slippery slope to epiphenomenalism. Then we discussed what is at stake if this is conceded.

We then talked about the possibility of identifying particular phenomenal experiences as certain brain states. If this is possible, we debated whether it would be the case that the brain state in question would be, say, the experience of red, even if it were unaccompanied by the corresponding phenomenal experience. Again, we discussed what the crucial **intension** of a concept should be, and debated whether it is just a matter of bashing intuitions if one does not accept the initial idea that the phenomenal experience of having a pain is something that needs to be explained.

Finally, Josh made a point about the possibility of himself not existing, which he has since posted to all.

Anthony and Rachael

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 24 12:17:24 1999
Date: Wed, 24 Feb 1999 13:05:41 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Rachael J Parkinson <rachaelp@U.ARIZONA.EDU>
Subject: Re: Kripke and the primary intension of "pain"
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I have a question regarding Kripke's characterization of heat on pg 133 and how it relates to pain. Kripke says, "Here heat is something which we have identified (and fixed the reference of its name) by its giving a certain sensation."(131) This would seem to pick out the primary intension of heat. But, Kripke points out, we have discovered a posteriori that heat is just molecular motion. Given that heat is just

molecular motion, it must be molecular motion in all counterfactual worlds. The primary intension of heat seems to be contingent given what we know about the secondary intension. So if someone were to say that light produces in them the same thing we feel when we feel heat, we would not say that light is heat but that the stream of photons produces a *sensation* of heat.

If we were to discover a posteriori that pain is certain C fibers firing in the brain, it seems like that would become the secondary intension of pain, the way heat is molecular motion is the secondary intension of heat. If this was the secondary intension of pain then it would fix pain to be "the firing of certain C fibers" in all counterfactual worlds. Thus, if someone were to say, light produces in me the same sensation that you feel when you feel pain we would not say that light is pain but that the stream of photons produces a *sensation* of pain.

If one was to argue that the *only* difference between heat and pain is that we have already discovered that heat is explainable in terms of molecular motion while we have not yet discovered that pain is explainable in terms of the firing of C fibers, it seems like the cases are relevantly similar. What do you think?
-Rachael

From owner-modality@LISTSERV.ARIZONA.EDU Wed Feb 24 20:12:50 1999
Date: Wed, 24 Feb 1999 21:10:10 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: Re: Kripke and the primary intension of "pain"
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on Rachael's note:

Kripke urges that, once we have identified heat as molecular motion, we have fixed the reference of 'heat'. As such, 'heat' would still refer to molecular motion if it happened that our nerve endings were different such that we actually had the sensation of cold when exposed to increased rates of molecular motion. I am not sure that I agree with this intuition-- it seems that Kripke is assuming the primacy of secondary intensions in this case. Suppose that one miraculously were to wake up in a world that is mostly like this one. In fact, it is, in many ways, indistinguishable from this one. The sun on one's skin feels warm and snow feels cold. Suppose that some time after relocating to this world one were to discover that heat actually corresponds to reduced molecular motion. The laws of physics in this world, however, are such that things behave exactly as they do in this world. It is not clear to me that one would, in this case, say, "Oh, I guess that what I thought was heat is not actually" upon discovering that heat is not more rapid molecular motion. It seems that this is a case where it is beneficial to think of the PI and the SI and consider which is more important to the concept of 'heat'.

I suppose that, in the case of heat, one could ignore the phenomenal component and say that heat really is only increased molecular motion. As I hope my example above shows, however, this does not seem entirely intuitive. I am not sure, in any case, that the experience of pain is a close analog to this one. Suppose that the phenomenal experience of having a particular pain is found to correspond to a particular brain state. Jane, a cutting edge neuroscientist, is satisfied that this is the case. But, suppose that Jane miraculously finds herself in a world slightly different from this one. Jane thinks she's still on earth, but, upon going to the lab and stimulating her brain in such a way that she is in the

state that she had identified as 'pain', she finds that her phenomenal experience is actually quite pleasant. It does not seem that this is a world in which pain is actually pleasant. Rather, it seems that Jane would admit to having misidentified pain as being a particular brain state

> I have a question regarding Kripke's characterization of heat on pg 133
> and how it relates to pain. Kripke says, "Here heat is something which we
> have identified (and fixed the reference of its name) by its giving a
> certain sensation." (131) This would seem to pick out the primary
> intension of heat. But, Kripke points out, we have discovered a
> posteriori that heat is just molecular motion. Given that heat is just
> molecular motion, it must be molecular motion in all counterfactual
> worlds. The primary intension of heat seems to be contingent given what
> we know about the secondary intension. So if someone were to say that
> light produces in them the same thing we feel when we feel heat, we would
> not say that light is heat but that the stream of photons produces a
> *sensation* of heat.
>
> If we were to discover a posteriori that pain is certain C fibers firing
> in the brain, it seems like that would become the secondary intension of
> pain, the way heat is molecular motion is the secondary intension of heat.
> If this was the secondary intension of pain then it would fix pain to be
> "the firing of certain C fibers" in all counterfactual worlds. Thus, if
> someone were to say, light produces in me the same sensation that you
> feel when you feel pain we would not say that light is pain but that the
> stream of photons produces a *sensation* of pain.
>
> If one was to argue that the *only* difference between heat and pain is
> that we have already discovered that heat is explainable in terms of
> molecular motion while we have not yet discovered that pain is explainable
> in terms of the firing of C fibers, it seems like the cases are relevantly
> similar. What do you think?
> -Rachael
>

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 26 23:28:07 1999

Date: Fri, 26 Feb 1999 23:27:54 -0800

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: Kripke and the primary intension of "pain"

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

In reply to Rachael and Anthony.

Rachael writes:

>If we were to discover a posteriori that pain is certain C fibers firing
>in the brain, it seems like that would become the secondary intension of
>pain, the way heat is molecular motion is the secondary intension of heat.
>If this was the secondary intension of pain then it would fix pain to be
>"the firing of certain C fibers" in all counterfactual worlds. Thus, if
>someone were to say, light produces in me the same sensation that you
>feel when you feel pain we would not say that light is pain but that the
>stream of photons produces a *sensation* of pain.
>
> If one was to argue that the *only* difference between heat and pain is
>that we have already discovered that heat is explainable in terms of

>molecular motion while we have not yet discovered that pain is explainable
>in terms of the firing of C fibers, it seems like the cases are relevantly
>similar. What do you think?

That's an interesting point. I think you're right that **if** we could wholly explain pain in terms of C-fibres, then we might well identify pain (in secondary intension) with C-fibres. The relevant questions are (1) How likely is this explanation, (2) What would follow from the identification?

Re (1), I think Kripke is more or less assuming the epistemic gap between physical states and phenomenal states, and is seeing what follows from that. He thinks it's conceivable (or "seems possible") that one could have an arbitrary physical state without a pain, for example. If so, then there already may be a principled explanatory gap between physical and phenomenal. To close that gap, presumably one would have to argue that there isn't really an epistemic gap after all -- either by arguing that there is an a priori entailment from physical to phenomenal after all, or by arguing that such an a priori entailment isn't required for reductive explanation.

With heat, on the other hand, it's not clear that there is the same sort of epistemic gap. If we conceive of heat as "the cause of heat sensations", and demonstrate that molecular property M is the cause of heat sensations, then we've arguably explained heat in terms of M. Well, maybe not quite, as heat sensations themselves haven't been explained, but we set that aside as a problem for later. Leaving that aside, we can say that M explains heat modulo consciousness (i.e. modulo problems with explaining heat sensations). Similarly, one has an a priori entailment from the physical story to the heat story modulo heat sensations -- no epistemic gap, except for the sensation gap. When it comes to heat, we don't let the sensation gap bother us, as we're concerned with the external phenomenon. But with pain itself (or pain qualia), it's the very sensation we're concerned with, so we can't set it aside in this way. So there is still the apparent epistemic gap.

Of course one could try to argue that there is no epistemic gap, presumably by rebutting the epistemic arguments we considered a few weeks ago. I take it that Kripke thinks that will be an uphill battle, though. In effect, he is here more concerned with the opponent who accepts that there is an epistemic gap but denies that there is an ontological gap. (Arguably the early identity theorists were like this, and many or most contemporary materialists are like this. That's the position I'm calling type-B materialism.) For that opponent, presumably things won't be so easy.

Incidentally we can put the points about epistemic gap and explanation themselves in terms of the 2-D framework. When it comes to explanation phenomenon P, it's arguably the primary intension of "P" that's central to characterizing what needs to be explained. To explain heat, for example, we essentially have to explain what causes heat sensations. To explain water, we explain watery stuff. And so on. The SI doesn't emerge until post-explanation. So something similar applies to consciousness. An explanation will need to entail the PI facts about consciousness; or, we'll need an epistemically transparent (a priori) entailment from physical to phenomenal. We seem to get that sort of entailment more or less when a concept's PI is functional or structural (e.g. the heat case is mostly functional, and the water case is functional/structural). But if the argument re conceptual analysis is right, it's not clear that there is a

functional/structural analysis of the PI of phenomenal concepts. If so, the epistemic gap looms, and explanation has a hard time getting off the ground.

(2) Some think that explanation isn't required for identity, and that maybe we could postulate an identification of pain with C-fibres (and of phenomenal with physical states more generally) without a prior explanation. Maybe correlation would give could grounds for the identity, for example. And some philosophers (e.g. Block, Papineau) have argued that identities don't need explanation. If we did this, we could identify pain in SI with C-fibres, and so on. But even here, I think there would be problems to overcome.

Essentially, even if we can pin down the SI of "pain" (etc.) in physical terms, there are still problems with the PI. One can argue that the very conceivability of zombies and the like (i.e. their broad epistemic possibility, in the considered-as-actual or PI sense) suggests that there is a world in the vicinity (by the 2-D analysis). If the materialist is right about the SI of phenomenal concepts, then this world shouldn't be described as one without pain (when considered as counterfactual), since the SI of "pain" will be physically specifiable and so will apply to this world. But it will still be a world in which the PI of "pain" isn't present. So there is *something* the world will lack -- essentially the property corresponding to the PI of "pain". We might think of this as the "mode of presentation" of pain. Even if we identify the *referent* of "pain" with a physical state, there is still the "sense" (or the reference-fixing property or the PI) to worry about -- the way that physical state is presented to us, if you like. And if the epistemic intuitions and the 2-D framework are correct, that will correspond to a nonphysical property in its own right.

The moral of all this is familiar, I think. To save materialism, one has to either deny the epistemic intuitions, or deny the 2-D analysis of a posteriori necessity. Either strategy remains open to some degree, but a materialist has to do a lot of hard work in biting the bullet on either of them.

Anthony writes:

>Kripke urges that, once we have identified heat as molecular motion, we
>have fixed the reference of 'heat'. As such, 'heat' would still refer to
>molecular motion if it happened that our nerve endings were different such
>that we actually had the sensation of cold when exposed to increased rates
>of molecular motion. I am not sure that I agree with this intuition-- it
>seems that Kripke is assuming the primacy of secondary intensions in this
>case. Suppose that one miraculously were to wake up in a world that is
>mostly like this one. In fact, it is, in many ways, indistinguishable from
>this one. The sun on ones skin feels warm and snow feels cold. Suppose
>that some time after relocating to this world one were to discover that
>heat actually corresponds to reduced molecular motion. The laws of physics
>in this world, however, are such that things behave exactly as they do in
>this world. It is not clear to me that one would, in this case, say, "Oh, I
>guess that what I thought was heat is not actually" upon discovering that
>heat is not more rapid molecular motion. It seems that this is a case
>where it is beneficial to think of the PI and the SI and consider which is
>more important to the concept of 'heat'.

Hmm, tricky. This case is complicated by your hypothesis that one actually wakes up in this world. In effect, you are considering the world as actual (wondering what we would say if we found ourselves in

it) rather than considering it as counterfactual. I think you're arguably right that if we found ourselves in this world, in which a different property R was causing heat sensations, we'd be tempted to say that R is heat. But that is arguably compatible with Kripke's framework, on which the PI of "heat" is "the cause of heat sensations" or some such.

(There are tricky issues here, regarding what one would say on first arrival in such a world vs. later on, etc. After a couple of years there, it seems pretty obvious one could correctly call R "heat". Does that apply even on first arrival? Just say one's nerve endings are tampered with so that temperatures below zero make one feel hot. Will one then correctly say that those temperatures are hot temperatures, or will you instead correctly say that cold temperatures are now making you feel hot. I can see both temptations, but I think there's a good case for the latter. That fits the idea that there's at least a time lag, and that the PI of "heat" is more like "what has typically causes heat sensations in me (and my community)", or some such.

But anyway, Kripke's central claim concerns the SI of "heat" rather than the PI. To get at that, we have to consider the world as counterfactual, rather than imagining waking up in it. From our perspective here, is that world (a world where -40 degree temperatures cause hot sensations) a world where -40 degrees constitutes heat. or a world where cold temperatures make people feel hot. Again tricky, but i can see Kripke's case for the latter. At least, the intuition for calling the low temperatures "hot" isn't as strong as when one thinks of waking up in the world (i.e., as when one considers the world as actual). I can still see a case for going both ways, though, which suggests that the SI of "heat" is arguably somewhat ambiguous.

>I am not sure, in any case, that the experience of pain is a
>close analog to this one. Suppose that the phenomenal experience of having
>a particular pain is found to correspond to a particular brain state.
>Jane, a cutting edge neuroscientist, is satisfied that this is the case.
>But, suppose that Jane miraculously finds herself in a world slightly
>different from this one. Jane thinks she's still on earth, but, upon going
>to the lab and stimulating her brain in such a way that she is in the
>state that she had identified as 'pain', she finds that her phenomenal
>experience is actually quite pleasant. It does not seem that this is a
>world in which pain is actually pleasant. Rather, it seems that Jane would
>admit to having misidentified pain as being a particular brain state

Interesting. Your intuition about the case seems pretty plausible. It seems unlikely that Jane would correctly call this a situation in which pain is pleasant. Rather, it would be a situation in which she isn't feeling pain. Of course this again is a point about PIs as described, not SIs. The moral of at least this scenario is that (if the situation is possible and if your intuition about describing it is correct) even if the SI of "pain" can be identified with a given brain state, the PI can't be. (In this world considered as actual, the brain state is present but pain isn't.) So there is still the issue about identifying the PI of "pain", as above.

One could also try to make a similar point about the SI of "pain". Let's take your Jane scenario and consider it as counterfactual instead of as actual. Should we (from our vantage point here) describe this as a situation in which Jane is having a pain (it just feels different), or one in which she is not having a pain at all. There is at least some pretty strong intuition that it should be the

latter, I think. If so (and if the situation is indeed possible), then the identification of the SI of "pain" with the brain state will be problematic, too.

In effect the relation between these two cases (considering a pain situation as actual and as counterfactual) parallels Kripke's observation re the semantics of pain. In effect, the fact that we have similar intuitions about saying whether a given scenario involves "pain", irrespective of whether it is considered as actual or as counterfactual, suggests that the PI and SI of "pain" (unlike "water", "heat", etc) are more or less the same as each other. I don't think that point is vital to making the anti-materialist case (the argument would still work even if PI and SI are different, as e.g. above and in the M&M notes), but it's an interesting point re the way the concept works.

Of course the materialist can try to reply by finding some brain state such that one couldn't imagine Jane having that brain state without pain. Maybe C-fibres are a particularly weak sort of state to pick, as we can just imagining wiring the brain so that C-fibres make her go "wow, this feels good", etc. So a better case for the materialist might be one where we have a full functionally-specified state, such that she'll always behave as if she's in pain, etc. Here the anti-materialist will still argue that it's conceivable that she have this without being in pain (cf. zombie cases), but at least it will be trickier for them. Again, the materialist either has to deny the epistemic intuitions (here by going for a functional analysis of the concept of "pain") or reject the 2-D framework (by arguing that the conceptually coherent scenario we are considering as actual doesn't correspond to a possible world at all).

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 26 10:33:53 1999

Date: Fri, 26 Feb 1999 11:31:34 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Timothy J Bayne <bayne@U.ARIZONA.EDU>

Subject: epiphenomenalism

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

I want to elaborate on the line I was developing on epiphenomenalism (E), and pretty much retract it. My reply to someone this week is to myself - hope that counts. AT least no one else can accuse me of misrepresenting them.

On tuesday Anthony asked why anyone would be tempted towards E. Isn't it obvious that the reason I drive out to watch the sunset is that I like the glorious red qualia that it gives me? Don't I like chocolate, and eat it, because of that chocolatey taste? Qualia seem to be causally efficacious.

Here's how I tried to pump the E intuition. Suppose that we accept the possibility of qualia inversion. Although you and I both like watching orange sunsets, the qualia that they produce in me are unlike the qualia they produce in you. Suppose that what it's like for me to see red is the same as what it's like for you to see green; and what it's like for you to taste ice-cream is what it's like for me to eat chocolate, and vice-versa. Inverted spectra intuitions are, I take it, rather easy to elicit, at least from those who are new to these issues. (Aged and battle-weary philosophers of mind seem not to share them.) I'm eating chocolate because I like its taste (note the reification of tastes). If it doesn't

have the same taste to you, but you still like it eating, the you are eating it for a different reason.

But how does inverted spectra provide an argument for E? My idea was that the very redness of red makes no difference to what it does, all that matters is that this experience be able to interact with my motor center in the right kind of way, and be the same kind of experience that I have when I see strawberries, fire-engines, and so on. The qualia that I have when I see red things can't be causally efficacious, since you experience different qualia, but act in the same type of way.

But this now seems to me to be a bad argument, or at least it's too quick. Two things can have the same causal effects even though they are different things. Thus, in certain conditions a white billiard ball will cause a window to break, and a red billiard ball will cause the window to break under the same conditions. But a red billiard ball has different properties from a white billiard ball. To rehabilitate the argument we need to talk about whether or not certain *properties* are causally efficacious. Rather than reify qualia, we need to talk about things that have the property of giving one a red appearance.

But even here the argument isn't obvious. It may be true that property x and property y have the same causal power in functional system F, but that means that x and y do have causal power within F - not that they have *no* causal power. (I guess the real intuition pump for epiphenomenalism is the zombie argument - but why would you think zombies possible attractive unless you were already an epiphenomenalist?)

Let's back up. The E claim is that the property of having qualia of a certain type are is not causally efficacious. I now think the motivation for it comes from the idea that the causal efficacy of a state/object only depends on its basic physical property, its 'syntax': weight, shape, magnetic force, whatever. (I think that both Scott and Thony said this, or things like it.) From this perspective, *most* properties are epiphenomenal. We say that the age of a vase causes it to be brittle, or valuable, but really these comments are just short-hand way of referring to the causal efficacy of the physical properties of states. The age of a vase causes it to be valuable = people believe that it is old, they believe that old things are valuable, both of these beliefs are physical states in the brain, and in virtue of their physical/syntactic properties, these states cause certain actions such as offering to pay a lot of money for the said vase.

Thus, we might say that age, or value, or qualia of certain kinds are causally efficacious properties in a *derivative* sense. Only synactical properties are causally efficacious in a primitive or basic sense. It seems that arguments for epiphenomenalism involve fairly large-scale metaphysical commitments re the nature of causation, and the structure of the mind. In short, the only motivation that I can see for E comes from the idea that all causation is bottom up. Talk of people, corporations, and countries, doing things is really just short-hand for the casual activity of basic material states. The debate over free will starts to creep in the backdoor here, so perhaps I'll shut up at this point.

Tim

Timothy J. Bayne
RM. 213 Social Science

Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Fri Feb 26 16:51:04 1999
Date: Fri, 26 Feb 1999 17:49:21 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik A Herman <erikh@U.ARIZONA.EDU>
Subject: TCM
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

There's one thing I need clarification on. One of the objections to naturalistic dualism was: There are properties essential to the physical constitution of the world that are not accessible to physical investigation...the zombie world seems physically identical while being physically different.

It seems that your simple response is that these extra properties are phenomenal. But I thought the objection was that there are might be extra PHYSICAL properties (that are presently not available to us).

Likewise, regarding your example of electrons that have hidden (protophenomenal) properties-- that we'd still call an electron that lacked these properties an electron. But again, wasn't the objection that there are extra physical properties (that missing, would presumably make it not an electron)?

The electron seems an appropriate example for I take this objection to be saying. Namely, that there are more (physical but wierd) properties that have yet to be understood with respect to the electron. Some of the physical properties that are presently being asserted are things like, "if we know this we can't possibly know that (and vice versa)" This isn't a phenomenal property but if it's a physical one it certainly isn't explanatory. This is the kind of stuff I would expect the objector to be referring to when talking about hidden physical properties-- the ones we don't "get", but not phenomenal ones.

So the question is, could these weird physical properties be the protophenomenal ones you propose?

-Erik H.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Feb 28 00:54:14 1999
Date: Sun, 28 Feb 1999 00:54:01 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: epiphenomenalism
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Tim and Erik H. raise interesting issues re epiphenomenalism and re the possibility of new intrinsic physical properties. These bear on the question of what the conclusion of the anti-materialist argument should be. I present it in the book as property dualism, and leave open just what sort. I now think that the conclusion of the argument should be a disjunction of three positions: epiphenomenalism,

interactionism, or "panprotopsychism". This issue isn't really core to the modal issues we're focusing on, but I'll say a bit about it anyway.

Epiphenomenalism is the idea that the phenomenal properties naturally supervene on physical properties and have no effect on them. (One-way psychophysical laws.) Interactionism is the idea that there are two-way psychophysical laws: the physical properties affect the phenomenal properties and vice versa. Panprotopsychism is the "Russellian" position mentioned a number of times in the book, on which the phenomenal is constituted by the unknown intrinsic properties of the physical.

I think the anti-materialist argument is actually neutral on these three, so the decision between them needs to be made on further grounds. I don't think there are yet grounds for a decisive preference in any direction. It seems to me, at least, that all have advantages and disadvantages. Personally, I rank panprotopsychism first (most elegant but also the least clear), epiphenomenalism second (arguably the most conservative, but inelegant), and interactionism third (would require adjusting our view of physics, but not out of the question). But I have days when each appeal.

Tim raises questions about epiphenomenalism, and Erik (in effect) about panprotopsychism. I'll say a bit in reply to Tim in the rest of this message and to Erik in the next.

Re epiphenomenalism first. Tim says quite rightly that the mere possibility that one can keep Q constant while varying P doesn't show that P is epiphenomenal with respect to Q. It might be that P1 and P2 can both cause Q in different circumstances. So if there is an argument for epiphenomenalism, that's not it.

I take it that the argument for epiphenomenalism is more indirect. It proceeds via an argument for property dualism, plus considerations about the causal closure of the physical. If qualia are nonphysical, and if there is a physical sufficient cause of any physical effect, then it seems that qualia are redundant. I take it that few people have the view that qualia are redundant as a premise, but one can see how one could be led to it as a conclusion. E.g., if one has accepted an anti-materialist argument and wants to hold on to the causal closure of the physical.

Just say one has causal closure and property dualism. Then presumably for behavior B, there will be a complete physical explanation P. In our world, there are qualia Q around too. But in other worlds, one has behavior B caused by P with different qualia Q' around, or with no qualia around at all. Even so, it seems that P is sufficient for a complete causal explanation of B -- it's not as if behavior is produced by a different mechanism in each case (as was the case in the causal explanations mentioned earlier). If so, then it seems that Q is causally redundant in the actual world. Barring an odd kind of overdetermination, Q seems to come out epiphenomenal.

Ways out of this: (1) Reject causal closure (go for interactionism). (2) Reject the modal intuitions (deny the possibility of inversions or zombies). (3) Accept overdetermination (both the physical processes and the qualia cause the behavior, redundantly). (4) Go for the Russellian/panproto view on which qualia are still causally relevant by virtue of their status as the categorical basis of physical dispositions.

The fourth would be my own preferred strategy here, as discussed around pp. 153-55 of the book. I think (3) is conceptually problematic, and obviously I don't accept (2). But all are at least respectable alternatives to consider.

I do think epiphenomenalism is a respectable alternative too. We have very strong intuitions of causal relevance for qualia, but those can be explained on Humean grounds: what we really know is that there is a constant regularity between qualia and behavior, from which it is natural to infer causation. But all our evidence is compatible with epiphenomenalism. So if there is reason to reject epiphenomenalism, it has to be based on further grounds, I think.

A couple of comments re Tim:

>causal power. (I guess the real intuition pump for epiphenomenalism is the
>zombie argument - but why would you think zombies possible attractive
>unless you were already an epiphenomenalist?)

I've heard this said before, but personally I think it gets the order wrong. Hardly anyone is "already an epiphenomenalist" in this context, but plenty of people find zombies conceivable. Maybe there is an argument from that **to** epiphenomenalism, although that is contestable. But in any case plenty of nonepiphenomenalists seem to accept the conceivability of zombies -- e.g. all the type-B materialists (Tye, Loar, Levine, Hill, Yablo, etc), and quite a few interactionists and others. More generally, I think one can assess the epistemic arguments against materialism prior to one's commitment on the metaphysical issue re epiphenomenalism.

>Let's back up. The E claim is that the property of having qualia of a
>certain type are is not causally efficacious. I now think the motivation
>for it comes from the idea that the causal efficacy of a state/object only
>depends on its basic physical property, its 'syntax': weight, shape,
>magnetic force, whatever. (I think that both Scott and Thony said this, or
>things like it.) From this perspective, **most** properties are
>epiphenomenal. We say that the age of a vase causes it to be brittle, or
>valuable, but really these comments are just short-hand way of referring
>to the causal efficacy of the physical properties of states. The age of a
>vase causes it to be valuable = people believe that it is old, they
>believe that old things are valuable, both of these beliefs are physical
>states in the brain, and in virtue of their physical/syntactic properties,
>these states cause certain actions such as offering to pay a lot of money
>for the said vase.

I'm not sure that this is getting at the real crux of the pro-epiphenomenalist argument in the qualia case. The argument you're giving will apply for **any** high-level property, including those which are clearly logically supervenient. I think such arguments will generally be suspect, as there's no reason to think that both micro properties and logically supervenient macro properties can't simultaneously be causally relevant (as e.g. with molecules and billiard balls being simultaneously causally relevant). Of course there is more to say there.

But in any case the special force of pro-epi considerations in the qualia case derives from the apparent failure of logical supervenience. That doesn't hold in those other cases, which have no apparent epistemic gap on close examination. But there is at least an apparent gap in the qualia case, which can be extended into an argument against logical supervenience and against materialism, etc, until (under certain assumptions) one gets to epiphenomenalism. Of

course there are contestable steps there. But take away the epistemic arguments, and I think the distinctiveness of the considerations re epiphenomenalism here are removed; and conversely, it's the apparent epistemic gap re qualia that gives pro-epi considerations here their distinctive force.

>In short, the only motivation that I can see for E comes from
>the idea that all causation is bottom up. Talk of people, corporations,
>and countries, doing things is really just short-hand for the casual
>activity of basic material states. The debate over free will starts to
>creep in the backdoor here, so perhaps I'll shut up at this point.

Well, maybe that is the only motivation for a *general* epiphenomenalist thesis, re arbitrary high-level properties. I think that motivation would be suspect, though it's an interesting issue in its own right. But in any case I think the motivation for epiphenomenalism re qualia rests on distinct considerations.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Feb 28 01:36:23 1999

Date: Sun, 28 Feb 1999 01:36:09 -0800

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: TCM (panprotopsychism)

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Now to Erik H.'s messages and the issues re "panprotopsychism". Erik wrote:

>There's one thing I need clarification on. One of the objections to
>naturalistic dualism was: There are properties essential to the physical
>constitution of the world that are not accessible to physical
>investigation...the zombie world seems physically identical while being
>physically different.

>

>It seems that your simple response is that these extra properties are
>phenomenal. But I thought the objection was that there are might be extra
>PHYSICAL properties (that are presently not available to us).

There are a few issues here. First we need to get clear on the nature of the new "hidden" properties. Perhaps this talk of "new" physical properties might suggest that they are just properties that will be invoked by a completed physics that we haven't yet discovered but eventually will, by scientific investigation and the like. Or perhaps that will never be discovered, but still have the same sort of general shape as existing properties. That would in effect be the "appeal to new physics" strategy considered in Chapter 3 of the book, which I argue there doesn't work. The trouble is that the new properties will still be introduced in a structural/dispositional way, and there will be no a priori entailment from these structural/dispositional properties to phenomenal properties, at least if the general epistemic considerations re qualia are valid. Those considerations seem to apply to any structural/dispositional properties, even as yet undiscovered one.

The move I'm suggesting here is a more radical one. Maybe in imagining a "physically identical" world, we are imagining a world that is the same in all structural/dispositional respects, but not the same in all *intrinsic* respects. That's to say, the "accessible"

properties that one holds constant are the structural/dispositional ones, but maybe there are further "hidden" intrinsic properties that are crucial to the physical world. And maybe those intrinsic properties entail the existence and nature of qualia.

So we now at least have some sort of property dualism, i.e. a dualism of "accessible" dispositional properties (I'll leave out "structural" from now on) and of "hidden" intrinsic properties. One might argue that this is out of the frying pan straight back into the fire, as the new protophenomenal properties seem to be epiphenomenal to the dispositions we know and care about. After all, the structural/dispositional properties seem to make up a seamless causally closed network. If the hidden intrinsic properties are indeed epiphenomenal with respect to the others, then it seems we at most have a terminological variant of the original property dualism (we're calling the new fundamental properties "hidden physical properties", but it doesn't really change the shape of the view). But I think there is one way of looking at this view so it at least has an interestingly different shape.

The interestingly different view is the one that locates the "hidden" intrinsic properties as the categorical bases of the dispositional properties. It's often held that dispositions need categorical basis -- where one has a disposition to cause Y, one needs some underlying intrinsic property X that actually does the causing of Y. E.g., a glass's fragility is a disposition, and its categorical basis is the molecular structure of the glass or some such.

Arguably, that applies all the way down to microphysics. Russell argued that physics tells us all about the *structure* of causal relations in the physical world, but it doesn't tell us what ultimately does the causing. Mass, for example, is treated by physics as a sort of disposition (the disposition to resist acceleration, be attracted by gravity, etc), but physics doesn't say anything about the basis of those dispositions -- what mass is "in itself". And more generally, one might argue that matter needs to have some intrinsic properties, not just dispositions, but physics is silent about that intrinsic nature.

So the interesting hypothesis is that we *need* certain "hidden" properties of the physical world to serve as the intrinsic basis of physical dispositions. These intrinsic properties will be what underlies the mass disposition that physics talks about, the charge dispositions, etc. Physical observation doesn't tell us anything about these properties, as we know about them only by their effects, and as far as the effects are concerned, arguably, any intrinsic basis will do. The hypothesis that physics needs hidden intrinsic properties is controversial, but many philosophers have found it attractive.

And note: *this* sort of hidden physical property is not epiphenomenal with respect to physical dispositions. The hidden properties here serve as the basis of physical dispositions, and arguably it is these that ultimately do all the causing. In fact the world might be seen as ultimately consisting in a bunch of causal and other nomic relations between a whole lot of intrinsic properties, which reveal themselves to us only structurally. So the hidden properties are at the very basis of the causal network.

So, now we have an interesting possibility re qualia. Maybe it's the case that the structural/dispositional physical properties don't entail qualia, but perhaps the hidden intrinsic properties do? After

all, we have no idea what those properties are like. Maybe they are themselves phenomenal properties, or some other more fundamental properties that entail phenomenal properties -- what I call protophenomenal properties. If so, then when imagining a zombie world, we're imagining a world with the same dispositional properties but different (or no) hidden intrinsic properties. So the conceivability of zombies is compatible with entailment by dispositional plus intrinsic physical properties. The cost of course is the presence of fundamental phenomenal or protophenomenal properties at the very fundamental level of the physical world. This is the position that I call "panprotopsychism". (In the book, "Russellian monism" or some such.)

In a way you can see this view as motivated by the presence of two metaphysical problems simultaneously: (1) What are the unknown hidden intrinsic properties of the physical world?; (2) How can we locate intrinsic phenomenal properties vis-a-vis the physical world? On having things put this way, it's natural to try to solve both at once: if the hidden intrinsic properties are themselves qualia (or protoqualia), then we both have a hypothesis about their nature and an integrated location for qualia in the physical world, as the categorical basis of physical dispositions.

Of course the view is highly speculative and has a number of problems it needs to face up to. But it has clear attractions also, not least the promise of an integrated, relatively monistic world view.

Anyway, with all this in hand, I'll get back to Erik's questions. Erik asked first, why can't the new properties just be ordinary physical properties, rather than phenomenal or protophenomenal properties? I think the answer is: first, the properties need to be intrinsic rather than dispositional, to avoid the usual problems with entailment. Second, the properties need to serve as the categorical basis of physical dispositions, to avoid epiphenomenalism. And third and most important, in order to deal with the qualia problem, the new properties have to either themselves be phenomenal properties, or they have to collectively constitute phenomenal properties -- i.e. they have to be phenomenal or protophenomenal. (Where protophenomenal is just a label for novel fundamental properties that are not themselves phenomenal but collectively constitute phenomenal properties.)

If the properties don't have this intimate relation to phenomenal properties, we'll still have the usual problems with zombies and the like. I.e. we'll still have the coherence of zombie worlds with the same accessible and hidden properties but no qualia; one could know all about the accessible and hidden properties without knowing about qualia; etc. But if they are phenomenal or protophenomenal, then it won't be coherently conceivable to have them instantiated without qualia, and in principle someone who knew all about the properties in question will be able to know about qualia. So these will need to be special sorts of properties.

>Likewise, regarding your example of electrons that have hidden
>(protophenomenal) properties-- that we'd still call an electron that
>lacked these properties an electron. But again, wasn't the objection that
>there are extra physical properties (that missing, would presumably make
>it not an electron)?

This gets at a question I've been avoiding until now. In what sense (if any) do these hidden properties count as *physical* properties, and so on what sense (if any) is the panprotopsychist view a variety

of physicalism? In the book, I mostly treat the view as if it is not a version of physicalism, but even there I say the issue is mostly terminological, and what matters is the shape of the view.

I think the issue comes down to: in what sense (if any) is a world with the same dispositional but different "hidden" properties "physically identical" to ours? In particular, does such a world truly count as containing mass, charge, electrons, etc? The neutral answer is: in the vicinity of mass, charge, electrons, etc, there are dispositional properties and intrinsic properties. The world in question has the dispositional properties but not the intrinsic properties.

But one might still press, and say -- but does it **really** contain mass, electrons, etc? This seems to be a largely terminological issue. We're considering a certain world -- the world with the same dispositional shape, different intrinsic properties -- as counterfactual, and evaluating the referent of "mass", "electron", etc, there. This comes down to an intuitions about the secondary intension of "mass", "electron", etc. I think it is highly plausible that the **primary** intension of these terms goes with certain dispositional properties (we fix reference to mass as "what plays the mass role", etc). But do the **secondary** intension go with the disposition or with the underlying basis?

In the book, I say that my intuition is that the world in question contains electrons, mass, etc. After all, the entities in question look like electrons, quack like electrons, interact like electrons, etc, and there's an intuition that that's enough to make them electrons. If so, the SI of "electron" (etc) goes with the disposition. But I can also see the opposing intuition. This would be the semi-Kripkean intuition that to be an electron (or mass), you have to have the "underlying nature" of electronhood (or mass), and here, that underlying nature is given by the intrinsic property. If so, something with a different intrinsic nature won't count as an electron, and the SI of "electron" goes with the intrinsic property, not the disposition.

If one takes the former view, then arguably the intrinsic properties are not physical properties, as they aren't the referent of paradigm physical terms like "mass", etc. And the zombie world in question will be correctly described as having mass (etc), so it will be a physical duplicate of our world. On this view, one might argue that the panprotopsychnist view is not a variety of physicalism, as the physical properties are structural/dispositional, not intrinsic.

If one takes the latter view, then arguably the intrinsic properties are paradigm physical properties, as they are the referent of paradigm physical terms like "mass", etc. And the zombie world in question won't be correctly described as having mass (etc), so it won't be a physical duplicate of our world. On this view, one might argue that the panprotopsychnist view is a variety of physicalism, where our fundamental physical properties turn out to be protophenomenal.

Ultimately (as I think I say in the book) I think the issue here is largely terminological, and nothing much turns on the semantic intuitions. We can say that the panprotopsychnist view is in a sense physicalist, and in a sense property dualist. It certainly has a dualism of dispositional and intrinsic properties; the only question is whether to call them "physical". I can see the case for going both ways, but in any case the shape of the view is what matters. And I think the shape of the view is pretty clear.

If one takes the version of the view on which the intrinsic properties are physical properties, then arguably we have "saved" physicalism in some sense from the anti-materialist arguments. But I think it's clear that the view that results is still quite radical, with new fundamental protophenomenal properties responsible for the emergence of qualia, and in some ways more in the spirit of property dualism than physicalism. At the end of the day it's the view rather than the label that matters.

>So the question is, could these weird physical properties be the
>protophenomenal ones you propose?

Yes, that's exactly it. I intended to gesture toward that idea in the passage you're talking about, but maybe the discussion wasn't too clear. (The view first comes up in considering the physicalist "objection" on pp. 136-38. On p. 138, I mention the ensuing possibility that physical entities have an intrinsic protophenomenal nature as a view that I'll return to. This was in effect a pointer to the discussion on pp. 153-56, which is the main discussion of the panprotopsychoist view in the book, though it's also discussed in the context of an informational view on pp. 301-8.)

But anyway, the view that fundamental protophenomenal properties serve as the "hidden" intrinsic categorical basis of fundamental physical dispositions is a very important view, no matter how one classifies it. It's a little on the wild and crazy side, but personally I find it the most elegant resolution of the mind-body problem if it can be worked out. It's compatible with all the epistemic intuitions re qualia, and at the same time is compatible a causal role for qualia *and* with the causal closure of the physical (qualia don't so much "change" the physical causal network as lie at its very basis). And the view has more of a monist flavor than the epiphenomenalist and interactionist alternatives, without the "dangling" entities that these views have. So there's a lot to be said for it.

On the negative side, one has to solve the problem of just how all these fundamental protophenomenal properties in physical systems come to constitute the (unified, bounded, homogeneous, structured, coherent) phenomenal experience we know and love. (See pp. 305-8 for a little discussion of this.) That's a highly nontrivial problem, and it's not obvious that it's solvable. But it's not obvious that it's unsolvable, either. So it's a very interesting topic for further research.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Feb 28 08:17:49 1999
Date: Sun, 28 Feb 1999 09:15:27 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Concepts and Properties
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I asked Josh and Thony the following question last week but our discussion got cut off so I thought I'd open it up. My query is this: in big picture terms, what exactly is the disagreement between Kripke and Chalmers. Here's how I think it goes - but I'm not clear that I've got it right.

Kripke elicits a bunch of intuitions about whether we'd call certain

things certain names under certain conditions;; i.e. if we found out that x was different in structure, would we still call it x. Kripke, or at least Kripkeans, take the results of these intuition pumps to tell us something interesting about *the world*, or about *properties*. They tell us that certain properties (x, y) are metaphysically linked: although there are possible world in which x and y come apart, they are identical (?) in all the metaphysically accessible worlds. Chalmers, I take it, responds thus: "the intuition pumps are persuasive, but they don't show what you (Kripkeans) take them to show. They tell us something about *language*, about our *concepts* <x> and <y> , and nothing about the world or about the structure of properties (x, y). Kripke's analysis doesn't show that the structure of possibility is other than what we have thought, all it shows is that we were misdescribing it (TCM 134)."

Two questions: (1) Is this the right way to characterize the dispute? (2) Isn't the connection between concepts and properties more intimate than Chalmers' line allows? If a concept applies in one scenario and not in another, doesn't that imply that the two scenarios do not share all of the same properties? Don't we generally decide whether or not we are dealing with the same property in different cases by deciding whether or not the same concept applies?

If we accept the move from conceivability to logical possibility, then (I think) we are accepting that there is definitely some connection between concepts and properties: a coherent use of concepts implies that the state of affairs (the articulation of properties) picked out by the concepts is logically possible. No doubt there is much to say on how properties and concepts are related, but I'll leave this here.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Mon Mar 1 05:06:16 1999
Date: Mon, 1 Mar 1999 05:06:01 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Concepts and Properties
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Tim writes:

>I asked Josh and Thony the following question last week but our discussion
>got cut off so I thought I'd open it up. My query is this: in big picture
>terms, what exactly is the disagreement between Kripke and Chalmers.
>Here's how I think it goes - but I'm not clear that I've got it right.

Interesting question. It's not entirely clear to me that there is a clear disagreement. Kripke doesn't use the 2-D framework, but it's not clear that he says anything that is strictly incompatible with it. And we have seen that Kripke seems to endorse a certain link between

conceivability and possibility, suggesting that when one apparently conceives of a world, there is at least a possible world in the vicinity, though it may be misdescribed. That view can be translated into the 2-D framework too.

I suppose one central disagreement is that I am more sympathetic with descriptive views of reference than he is, and this shows up in my explicit use of the notion of a primary intension, something he stays away from. But one might argue that the notion of a primary intension is at least implicit in some of Kripke's discussion, as when he evaluates what the referent of our terms would be if we discovered certain epistemic possibilities to be actual.

I suppose that one difference is that Kripke sees epistemic matters such as apriority as **purely** epistemic in some sense and not modal, whereas I think it is often useful to understand them in modal terms. On my view, one can see epistemic possibilities as possible worlds in their own right, where our terms have a distinctive pattern of application to those worlds; and there is a corresponding notion of necessity, 1-necessity or primary necessity, which is in many ways just as important as the notion of subjunctive necessity (2-necessity) that Kripke concentrates on, and which has deeper ties to epistemic notions.

Probably in the end these are the most important difference -- a difference of emphasis on PI vs. SI considerations, a willingness to see some sort of quasi-descriptive component in meaning, and a willingness to understand epistemic notions in modal terms. But these don't immediately manifest themselves in first-order issues or doctrines. It is more a difference in emphasis, and in the lens through which one looks at the central issues.

Personally, I like to think of the 2-D framework as a good way of analyzing and making sense of what Kripke is up to in N&N, rather than as any sort of radical revision of the views there. Of course Kripke might see things differently.

>Kripke elicits a bunch of intuitions about whether we'd call certain
>things certain names under certain conditions;, i.e. if we found out that
>x was different in structure, would we still call it x. Kripke, or at
>least Kripkeans, take the results of these intuition pumps to tell us
>something interesting about **the world**, or about **properties**. They tell
>us that certain properties (x, y) are metaphysically linked: although
>there are possible world in which x and y come apart, they are identical
>(?) in all the metaphysically accessible worlds. Chalmers, I take it,
>responds thus: "the intuition pumps are persuasive, but they don't show
>what you (Kripkeans) take them to show. They tell us something about
>**language**, about our **concepts** <x> and <y>, and nothing about the world
>or about the structure of properties (x, y). Kripke's analysis doesn't
>show that the structure of possibility is other than what we have thought,
>all it shows is that we were misdescribing it (TCM 134)."

Well, I don't disagree with the claims you attribute to Kripke here. It's quite compatible with the 2-D framework to say that Hesperus is identical to Phosphorus in all possible worlds, and that the property of being hot is the property of having fast molecules in all possible worlds. So I can agree that the objects and properties in question are metaphysically linked here; or better, that there is just one object or property in each case, with two different terms picking it out.

Of course I think there is more to the story than this. In making the

above claims I am dealing with 2-possibility and secondary intensions. One can also invoke 1-possibility and primary intensions, and explain the a posteriority of "Hesperus is Phosphorus" in terms of the fact that its PI is false in some centered worlds, and so on. But to say this is not to take back any of the above.

It's true that I am also inclined to think that some of these issues are vague and depend on the terminology. For example, it's not obvious to me that the SI of "heat" really picks out the motion of molecules rather than the cause of heat sensations in all possible worlds. But again, this isn't to retract the claim about properties. It's just to say that the reference "heat" is somewhat ambiguous between two properties, the molecule property and the heat-sensation property (N.B. this isn't an issue about PI vs. SI; rather, it's an ambiguity merely in the SI). But like any ambiguity, this can easily be cleaned up or stipulated away, and after doing this, we'll have a perfectly good term, "heat" or "heat*" which works just the way Kripke says it does, referring to the molecule property, and picking out that very property in all worlds. We can then say that heat (or heat*) is identical to molecular motion in all worlds. So I don't think the point about vagueness and ambiguity is central here.

Deep down you are probably right that I am more deflationary about the metaphysics of properties and possibility here than Kripke is, but it's not obvious how to articulate that. Sure, the property of heat and the property of molecular motion are the same property. But the fact that "heat is molecular motion" is an a posteriori necessity here is largely due to our conceptual structure, with two distinct concepts that pick out the same property in different ways. Similarly, it's true that it's necessary that water is H₂O, and that fact is entirely independent of us. But the fact that "water is H₂O" is an a posteriori necessity simply arises from the coreference of the terms and the 2-D structure of the concepts. Even here, it's not entirely clear what Kripke would disagree with, so I'm not sure that I've really isolated a substantive disagreement.

As for the conceivability/possibility thesis you mention, Kripke shows every sign of agreeing with it, so it's not clear that this is a disagreement.

>Two questions: (1) Is this the right way to characterize the dispute? (2)
>Isn't the connection between concepts and properties more intimate than
>Chalmers' line allows? If a concept applies in one scenario and not in
>another, doesn't that imply that the two scenarios do not share all
>of the same properties? Don't we generally decide whether or not we are
>dealing with the same property in different cases by deciding whether or
>not the same concept applies?

Well, of course I think there are two ways in which concepts can apply to scenarios: the PI way and the SI way (considering the scenarios as actual and counterfactual). We can take these cases one at a time.

Taken in the SI way (considering scenarios as counterfactual): if a concept applies in one scenario but not in another, then the SI of the concept applies to one world and not to the other, so the worlds certainly have different properties (as SIs are properties in their own right). For example, the fact that the concept "water" applies to H₂O in the H₂O-world (considered as counterfactual) but not to XYZ in the XYZ-world (considered as counterfactual) goes along with the fact that they have different properties: H₂O has the property of being water and XYZ does not! It will also be the case that we can

determine whether we have same property X in two scenarios by seeing whether the concept "X" applies to those scenarios (considered as counterfactual). So there's nothing here with which I disagree.

Taken the PI way (considering scenarios as actual): if a concept applies in one scenario but not in another, then the PI of the concept applies to one world and not to the other, so the worlds certainly have different properties (as PIs are properties in their own right). For example, the fact that the concept "water" applies to H₂O in the H₂O-world (considered as actual) but not to H₂O in a world where H₂O is a coal-like solid (considered as actual) goes along with the fact that they have different properties: H₂O in the first world has the property of being watery stuff but in the second world does not! So your first claim will hold good even for PIs.

The second claim is a bit trickier. Can we decide whether or not we are dealing with the same property by seeing whether the same concept applies (to a scenario considered as actual)? Well, in a way yes, in a way no. The concept "water" will apply to first but not the second of the scenarios considered above. But arguably both scenarios contain the property of being water. If so, then the application of the concept "water" (in PI) is not a perfect guide to the presence of the property of being water (the property is present in the second H₂O-world, even though the PI of "water" doesn't apply there). On the other hand, the application of the concept "water" to these scenarios (considered as actual) *will* be a good guide to the presence of a certain property, the property determined by the PI of water, i.e. the property of being watery stuff. The concept "water" applies to the first but not the second scenario (considered as actual), and the property of being watery stuff is present in the first but not the second. So even here, the application of a concept is a good guide to the presence of a corresponding property -- it's just that in this case, it's the PI property that's relevant, not the SI property.

Let me know if I've misunderstood your claim here.

>If we accept the move from conceivability to logical possibility,
>then (I think) we are accepting that there is definitely some connection
>between concepts and properties: a coherent use of concepts implies that
>the state of affairs (the articulation of properties) picked out by the
>concepts is logically possible. No doubt there is much to say on how
>properties and concepts are related, but I'll leave this here.

Right. On my view, the connection between concepts and properties is particularly intimate, more intimate than some opponents will allow. (So I'm not sure why you say above that that connection may be more intimate than *I* allow.) In particular, I think that when one has two a priori distinct concepts, there are two distinct properties in the vicinity, namely the two reference-fixing properties, or PI properties. (If "A = B" is a posteriori, the PI of "A" is a distinct property from the PI of "B".) This issue comes up a bit in the PPR symposium we'll be looking at shortly.

Re moving from conceivability to possibility, you're right again. Of course I think there are two such moves in the vicinity: the move from 2-conceivability to 2-possibility and the move from 1-conceivability to 1-possibility. Either way there will be a link between concepts and properties. In fact, I think in both cases, the link between conceivability and possibility and the link between concepts and properties stand and fall together. To see this in a bit more detail:

Thesis 1: 1-conceivability implies 1-possibility.

Thesis 2: Distinct concepts yield distinct properties. I.e., if "A" and "B" are distinct concepts, in the sense that there are 1-conceivable scenarios to which "A" applies but "B" does not, then the PI of "A" and the PI of "B" are different properties, in the sense that they are not necessarily coextensive.

Claim: Thesis 1 $\leftrightarrow$ Thesis 2.

Proof:

Left-to-right: Assume 1-conceivability implies 1-possibility. Let "A" and "B" be distinct concepts. Then there are 1-conceivable scenarios to which "A" applies but "B" does not (i.e., "A $\neq$ B" is 1-conceivable). It follows that "A $\neq$ B" is 1-possible, and there are 1-possible scenarios to which "A" applies but "B" does not. So the PI of "A" is not necessarily coextensive with the PI of "B", so these are distinct properties.

Right-to-left: Assume that different concepts yield different properties. Let "A $\neq$ B" be 1-conceivable. Then "A" and "B" are distinct concepts, so yield distinct properties, so the PI of "A" is a property distinct (and not necessarily coextensive with) the PI of "B". So there is a world to which the PI of "A" applies but the PI of "B" does not. So "A $\neq$ B" is 1-possible.

Much the same goes for the corresponding claims regarding 2-conceivability, 2-possibility, and SIs. One needs a little more to generalize the right-to-left proof for conceivability-to-possibility links in cases other than identity statements, but it is not too difficult.

So, we can see that the sort of 2-D view to which I subscribe (and which I think is independently plausible) has a particularly strong link between concepts and properties (distinct PIs yield distinct properties, etc) and between conceivability and possibility (1-conceivability implies 1-possibility, etc), and that these two properties of the framework go together.

It seems, then, that it is an opponent who will have to resist this strong link between concepts and properties, at least if they want to resist the move from (ideal) conceivability to possibility. We'll see some instances of that strategy when we discuss the papers by Loar and by Hill & McLaughlin in the PPR symposium. Essentially, these people reject the 2-D framework as a complete account of a posteriori necessity, reject the central claim that an a posteriori necessity must have a contingent primary intension, and deny the strong link between concepts and properties and between conceivability and possibility. It seems that's what an opponent (at least a type-M materialist) needs to do, if they want to hold onto the epistemic intuitions but still hold on to materialism. That's something we'll be looking at in depth soon.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Mar 1 07:40:05 1999
x-sender: agillies@pop.u.arizona.edu
Date: Mon, 1 Mar 1999 08:53:38 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>

Subject: Re: Concepts and Properties
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Dave recently responded to Tim's question re "the big picture disagreement" between 2D views and those in N&N by (among other things) noting that:

>I suppose that one difference is that Kripke sees epistemic matters
>such as apriority as *purely* epistemic in some sense and not modal,
>whereas I think it is often useful to understand them in modal terms.
>On my view, one can see epistemic possibilities as possible worlds in
>their own right, where our terms have a distinctive pattern of
>application to those worlds; and there is a corresponding notion of
>necessity, 1-necessity or primary necessity, which is in many ways
>just as important as the notion of subjunctive necessity (2-necessity)
>that Kripke concentrates on, and which has deeper ties to epistemic
>notions.

There may be another way of getting at this point of tension between the two views. This relates to some of Josh's worries about worlds considered as actual. On the 2D view of things, considering a (centered) possible world as actual is straightforward enough, and there is a well-defined and well-behaved notion of necessity and possibility that corresponds to the application conditions of terms in worlds-considered-as-actual. One reason why this is easy to do on the 2D view is that worlds are considered purely qualitatively (see TCM, p.367n30). And so, a la Lewis, the actual world is privileged not because it is actual. but because we happen to inhabit it: other worlds are truthfully called "actual" by their inhabitants.

But Kripke doesn't seem to want to think of worlds in that way. He even scorns the literature on trans-world identity for being confused: these philosophers wouldn't have such difficulty if they didn't think that worlds could only be given qualitatively (p.76). Worlds, on the Kripkean view, are *always* counterfactual worlds---at least that seems to be the way Kripke talks: worlds are total ways the world *might have been* (p.18). So, it's no wonder he focuses on SI's---strictly speaking, he has nothing else to talk about.

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 2 00:37:01 1999
Date: Tue, 2 Mar 1999 00:36:41 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Concepts and Properties
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Thony writes:

>There may be another way of getting at this point of tension between the
>two views. This relates to some of Josh's worries about worlds
>considered as actual. On the 2D view of things, considering a (centered)
>possible world as actual is straightforward enough, and there is a
>well-defined and well-behaved notion of necessity and possibility that
>corresponds to the application conditions of terms in
>worlds-considered-as-actual. One reason why this is easy to do on the 2D

>view is that worlds are considered purely qualitatively (see TCM,
>p.367n30). And so, a la Lewis, the actual world is privileged not
>because it is actual. but because we happen to inhabit it: other worlds
>are truthfully called "actual" by their inhabitants.

Well, I wouldn't go so far as to say that the 2-D view is Lewis-like.
Lewis's view holds in effect that all worlds are equally "concrete",
and that what distinguishes the actual world is just that we are in
it. But it's quite compatible with the 2-D framework to say that only
the actual world is concrete while the others are abstract, and this
concreteness distinguishes it from the other worlds, not just the fact
that we are in it. Or to use other language, we can say that there is
a large space of (abstract) possibilities, but only one of the
possibilities is **realized**, namely this one.

Of course, one can consider any possible world as actual, but that's
not to say that the actual world isn't privileged. This world
is not merely considerable-as-actual -- it **is** actual!

>But Kripke doesn't seem to want to think of worlds in that way. He even
>scorns the literature on trans-world identity for being confused: these
>philosophers wouldn't have such difficulty if they didn't think that
>worlds could only be given qualitatively (p.76). Worlds, on the Kripkean
>view, are **always** counterfactual worlds---at least that seems to be the
>way Kripke talks: worlds are total ways the world **might have been**
>(p.18). So, it's no wonder he focuses on SI's---strictly speaking, he
>has nothing else to talk about.

Hmm, lots of interesting issues here. I suspect that the issue
vis-a-vis qualitative specification and transworld identity is to some
extent separable from the issue about considering as actual and as
counterfactual, though there are links.

Kripke says that possible worlds **needn't** be specified in mere
qualitative terms, though he also allows that they **can** be. So I can
if I like specify a counterfactual world in terms of the distribution
of qualities, but I can also specify it as one in which **Nixon** does
such-and-such, and so on. Kripke thinks this cleanly bypasses any
problems about "transworld identity". Of course there might still be
a problem concerning identity across worlds given qualitatively, since
Kripke at least allows the latter, but the fact that one needn't
always specify a world qualitatively means that there isn't a problem
about transworld identity **in general**.

As for worlds being considered as counterfactual, it's true that
insofar as Kripke explicitly talks about what is true in a world, he
is always considering those worlds as counterfactual, and thus in
effect invoking SI evaluation. He does seem to often **implicitly**
consider worlds as actual and evaluate the referent of our terms there
(as with considering the epistemic possibility that Godel stole the
proof, or thinking about what we'd say if we discovered that the
catlike things were demons); but he doesn't explicitly put this in
terms of evaluating terms in possible worlds. So at least as far as
his explicit framework goes, possible worlds are always considered as
counterfactual (though I'd argue that he is implicitly endorsing
something more).

(You're right that thinking of worlds as ways things might have been
goes along with considering them as counterfactual. One way of
bringing out the 2-D distinction is to say that worlds considered as
actual are ways the world **might be**, while worlds considered as
counterfactual are ways the world **might have been**. This goes along

with the ties to indicative and subjunctive mood in conditionals and elsewhere. Kripke always focuses on the latter, it seems.)

What's the relation between the two points: allowing us to stipulate worlds nonqualitatively, and always considering worlds as counterfactual? The former does not seem to imply the latter: someone might accept that we can stipulate counterfactual worlds qualitatively but also accept that we can consider worlds as actual (whether qualitatively or nonqualitatively). And the latter does not seem to imply the former: someone might always consider worlds as counterfactual but also always consider them qualitatively ("through a powerful telescope"), not allowing nonqualitative stipulation. So the two points seem independent to some degree.

Still, maybe there is some indirect link. One way to get at this is to ask whether it is possible, when considering worlds as actual, to stipulate them nonqualitatively as well as considering them qualitatively. To a certain degree this seems to be possible: for example, I can stipulate that I am talking about the epistemic possibility that Nixon was Kennedy's brother, rather than considering the world "given qualitatively" and trying to see whether it is an instance of the epistemic possibility that Nixon was Kennedy's brother. So to that extent Kripke's point holds even here. (Of course, in both cases one needs to be careful that one is not making an incoherent stipulation.)

Still, there's a difference here in that arguably the notion of transworld identity doesn't make as much sense when talking about worlds considered as actual. I can consider the epistemic possibility that Hesperus is not Phosphorus (that's pretty straightforward), but presumably that's not a world where **this very thing** (i.e. Hesperus/Phosphorus/Venus) is not **that very thing** (i.e. Hesperus/Phosphorus/Venus). Presumably that would be incoherent. So when we stipulate the epistemic possibility, one is not making a stipulation about the very things from the actual world.

Arguably, the notion of transworld identity does not really make sense when applied to worlds considered as actual, as opposed to worlds considered as possible. The reasons for this are subtle but come down at least in part to the failure of substitutivity in epistemic but not in modal contexts (and ultimately in indicative but not in subjunctive contexts). Epistemic possibilities are not "object involving" in quite the same way that subjunctive possibilities are. So **de re** (object-involving) possibility and necessity may not really be coherent here in this domain.

For example, it seems one can't quantify into 1-modal contexts: "it is 1-necessary that x is hot" isn't really coherent. E.g., it is 1-necessary that Hesperus is the evening star, but it isn't 1-necessary that Phosphorus is the evening star, even though both are names for the same thing; and there isn't really any other way to make sense of the claim without such names. So for the epistemic modality, any modal claims involving x seem to be relative to the label under which x is picked out. This doesn't apply to subjunctive modality: "it is 2-necessary that x is hot" makes sense, and any label for x will yield the same truth-value here, as long as it's a rigid designator.

(In a way, this is saying that Quine's critique of quantified modal logic and essentialism was right about epistemic modality, though not about subjunctive modality. So we have a "split decision" in the battle between the likes of Quine and the likes of Kripke. A point

like this is made John Burgess in a very nice recent paper in the 1997 Canadian Journal of Philosophy supplement on "Meaning and Reference".)

Anyway, nothing here contradicts anything that Kripke says, exactly. >From the 2-D perspective, one can say that he is focusing on the sort of possibility and necessity (namely subjunctive possibility and necessity) where de re possibility and necessity does make sense, and it's quite an achievement of his to make sense of it. At the same time, there are other notions of possibility and necessity out there (e.g. broadly epistemic possibility and necessity) where de re possibility and necessity doesn't make sense.

Of course it's no surprise that Kripke, who was such a major contributor to quantified modal logic, was particularly interested in investigating notions of modality for which quantification into modal contexts makes sense (which will be just those for which *de re* modality makes sense). So it's no surprise that he was led to focus on subjunctive modality, and on worlds considered as counterfactual. (More on this in the "tyranny of the subjunctive" material.)

But anyway, none of this implies that there's anything problematic or second class about epistemic modality and considering worlds as actual. Things here just work a different way. One can take on board all of Kripke's points about subjunctive modality, and still hold that there are other interesting modal notions and other corresponding ways of looking at possible worlds.

Or so it seems to me. But feel to elaborate on this interesting issue.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 2 00:01:44 1999
Date: Tue, 2 Mar 1999 00:16:22 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I have a question concerning the possibility of psychophysical laws, given the anti-materialist argument we've been discussing.

One of the straightforward problems that one adopts, if the non-(logical supervenience) of consciousness is granted, is the rather unpalatable possibility of epiphenomenalism with respect to conscious experience. Chalmers' remarks that this may not be so bad, and that, if the argument goes through, we may simply "have to get used to it". But I take it that Chalmers' (self-admitted!) speculation about the possibility of internal, protophenomenal features of an "expanded" physical framework, who's current postulates are all external, relationally specified properties (i.e., of the fundamental physics variety) is meant to provide a possible future development of a causal role for phenomenal properties. (Is this true?). I have two questions. First, if we have some set of internal (protophenomenal) properties related by psychophysical laws, and still the external, physical properties related by physical laws, then don't we just push the problem of interaction (and thus epiphenomenalism) back a step?

Second, if we have the psychophysical laws connecting internal to external properties--protophenomenal to physical--wouldn't this fly in the face of a) the non-(logical supervenience) of the phenomenal on the physical and

b) the assumption that the physical world is causally closed? All of this is to suggest that there doesn't seem to be much room for psychophysical laws, at least in so far as they may provide possible answers to the problem of epiphenomenalism. One response may be that considering psychophysical laws qua answers to epiphenomenalism is not necessary, anyway. But I take it that a "completed" scientific account of phenomenal properties in a theory of the mental would have something like this in mind.

(There are really two (or more) possibilities for the role of psychophysical laws relevant here. Qua answer to epiphenomenalism, the laws would be causal. A "weaker" interpretation would assign psychophysical laws the role of correlating internal (protophenomenal) to external (physical) states in a way that yielded information about regularities between the two. I suppose I'll need to think things through more for the weak case, but it seems that the same sort of objections stated above will present themselves.)

Let me know what you think.

Erik

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 2 14:37:06 1999
Date: Tue, 2 Mar 1999 12:57:29 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Angela J Burnette <aburnett@U.ARIZONA.EDU>
Subject: Re: epiphenomenalism
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I realize that what follows is actually last weeks topic, but here goes anyway...

In an attempt to avoid all out epiphenomenalism, it is suggested in TCM and in Metaphysics of Modality that there might be a causal role for the phenomenal if the intrinsic properties of microphysical stuff is actually phenomenal or protophenomenal, this is a possibility I take it because we understand this stuff only in virtue of its extrinsic properties, which leaves the intrinsic nature up for grabs...if the intrinsic properties were phenomenal or protophenomenal, then perhaps they could form the underlying basis for the physical realm, and therefore somehow be causally related, even though the physical system is causally closed...

I have two questions about how this might go, first, if protophenomenal properties are intrinsic, then what access will we ever have to them such that laws relating them might be developed? Also, how is it that "principles associating physical dispositions with (proto)phenomenal bases" is less problematic than laws relating phenomenal properties and

physical properties? I'm not sure how one can posit a "transformation operator, for example" given all that has gone before about the utter impossibility of cashing out phenomenal properties in functional or dispositional terms.

so, it seems as though epiphenomenalism cannot be avoided without undermining the anti-materialist arguments...

angela

From owner-modality@LISTSERV.ARIZONA.EDU Thu Mar 4 04:32:01 1999
Date: Thu, 4 Mar 1999 04:31:49 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: TCM (panprotopsychism)
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Both Erik L. and Angela raise questions about panprotopsychism, vis-a-vis epiphenomenalism.

Erik writes:

>I take it that Chalmers' (self-admitted!) speculation about the possibility
>of internal, protophenomenal features of an "expanded" physical framework,
>who's current postulates are all external, relationally specified
>properties (i.e., of the fundamental physics variety) is meant to provide a
>possible future development of a causal role for phenomenal properties.
>(Is this true?). I have two questions. First, if we have some set of
>internal (protophenomenal) properties related by psychophysical laws, and
>still the external, physical properties related by physical laws, then
>don't we just push the problem of interaction (and thus epiphenomenalism)
>back a step?
>
>Second, if we have the psychophysical laws connecting internal to external
>properties--protophenomenal to physical--wouldn't this fly in the face of
>a) the non-(logical supervenience) of the phenomenal on the physical and
>b) the assumption that the physical world is causally closed? All of this
>is to suggest that there doesn't seem to be much room for psychophysical
>laws, at least in so far as they may provide possible answers to the
>problem of epiphenomenalism. One response may be that considering
>psychophysical laws qua answers to epiphenomenalism is not necessary,
>anyway. But I take it that a "completed" scientific account of phenomenal
>properties in a theory of the mental would have something like this in
>mind.

Well, one has to get the shape of the view straight. On the "panprotopsychist" view, one has still has a causally closed physics. In fact physical laws have just the same shape as before. We have just "colored in" physics by specifying the categorical basis of the physical dispositions.

Think of it this way. Take physics at face value as true. Even so, one arguably realizes that physics only tells us about the "structure" in relations between things, and tells us about things only by their effects. What is a particle? What is mass? Physics doesn't say. We just know about the effects of these things.

More technically, we just know about the *dispositions* in the vicinity of charge and mass and electrons. We know about them solely in terms of their characteristic effects, which are well spelled-out

by physics. But arguably, wherever one has a disposition, one needs a categorical basis. That's just to say, whenever one has a causal role, one needs something to do the causing! That thing will be the "intrinsic" nature of mass, or charge, or an electron. Physics doesn't tell us about it -- it is in effect hidden from sight, in a sort of Kantian way -- but arguably we still have reason to believe it exists.

On this view, every physical property -- mass, charge, etc -- has a categorical bases, or an intrinsic nature. A fundamental particle will have an intrinsic nature given by the nature of its charge and mass, etc. The world itself is a giant network of causal interactions among these intrinsic properties and the particles that have them. The interactions are structured just physics says -- the laws of physics specify the dynamics perfectly. It's just that physics tells us only about the structural part, and we can't forget the intrinsic part too.

Panprotopsyichism is the view that these intrinsic properties are phenomenal or protophenomenal. So the physical world consists ultimately in interactions among phenomenal or protpphenomenal properties. It doesn't look that way to us, because we see the properties only from the "outside" (except in our own case!), in terms of the structure of their interactions. But on this view, it's that way all the same. After all, it's no surprise that we don't know the nature of the intrinsic properties of the physical.

Now to Erik's questions:

(1) Doesn't this just push the problem of interaction and epiphenomenalism back a step? I'd like to think not. The beauty of doing things this way is that we *know* that the intrinsic properties are causally relevant. They are the things that do the ultimate causing! E.g., when mass causes another mass to be attracted, what is really happening is that there is a causal relation between their intrinsic bases, i.e. (on the panprotopsyichist view) the relevant protophenomenal properties. The world is ultimately a network of causal interactions among just these properties! So they are causally relevant from the "bottom up".

(2) What about causal closure? Another beauty of this view is that one can hold onto causal closure. We haven't introduced new causal "gaps" from phenomenal propeties to fill. Rather, they just play the role of supporting the causation that was there all along. Interactions among particles etc can be just as they were before; it's just that we now know that the ultimate nature of those particles is protophenomenal.

(3) What about psychophysical laws? These have a slightly different shape on this view. They don't connect physical properties to phenomenal properties "dangling" outside the system. Rather, they will in effect connect "dispositional" properties (e.g. the characteristic causal role of mass) with the "intrinsic" properties that underlie the disposition (and that are ultimately doing the causing). In effect, they are connecting the "outside" of physical entities to the "inside".

Arguably, the truly fundamental laws in this world are laws connecting the basic intrinsic properties to each other. e.g. protophenomenal property 1 (the mass property) and protophenomenal property 2 (the charge property) interact with each other in such-and-such a way. The "dispositional" physical properties just come from our viewing all

this from the outside in terms of structure, and the "psychophysical laws" from our trying to fill in the dispositions with intrinsic properties again. But at least from our perspective, it's useful to see these psychophysical laws here, as the dispositional stuff is what we're familiar with (from physics etc), and the intrinsic stuff is all a bit of an unknown. So one can hope that a final science of consciousness might make some progress in connecting the two, and in particular telling us just which intrinsic properties play which dispositions.

(4) Logical supervenience on the physical? As I said in the earlier message, there's a sense in which this view sees phenomenal or protophenomenal properties as part of the physical world. After all, they are the categorical bases of physical dispositions, and arguably are the "ultimate nature" of the physical. But still, they don't logically supervene on the physical properties we know and love, because those are dispositional properties, and intrinsic properties don't supervene on dispositions. It would be logically possible to have the same dispositions but a different intrinsic basis, etc. But here, the failure of supervenience isn't due to the intrinsic properties being "outside" the network. It's because it's the intrinsic properties that "carry" the dispositions, serve as their "insides", and ultimately do the causing.

Angela writes:

>I have two questions about how this might go, first, if protophenomenal
>properties are intrinsic, then what access will we ever have to them such
>that laws relating them might be developed? Also, how is it that
>"principles associating physical dispositions with (proto)phenomenal
>bases" is less problematic than laws relating phenomenal properties and
>physical properties? I'm not sure how one can posit a "transformation
>operator, for example" given all that has gone before about the utter
>impossibility of cashing out phenomenal properties in functional or
>dispositional terms.
>
>so, it seems as though epiphenomenalism cannot be avoided without
>undermining the anti-materialist arguments...
>
>angela

Good questions. First of all, let me say that I'm not suggesting that the problems you mention are any easier on a panprotopsychoist view than on an epiphenomenalist view. But I don't think that they're necessarily harder, either.

E.g., for your first problem about "access", obviously we don't have much access to epiphenomenal phenomenal properties, except in our own case. And indeed one can argue that almost any view will have some sort of epistemic barrier of this sort, even a type-B materialist view: one can't monitor qualia in other systems directly, and there are always multiple epistemic possibilities. So the panprotopsychoist view isn't obviously worse off here.

Similarly for the second problem. Our "transformation operator" isn't going to try to reduce phenomenal properties to dispositional ones. We are just going to try to find a law connecting the two (laws that connect distinct things, of course). That will be hard for an epiphenomenalist, but there's no reason why it should be impossible in principle. Hopefully a panprotopsychoist can do the same.

How do we find the laws? Well, first of all, by correlating and systematizing first-person data and its relation to third-person data in one's own case. And second, by gathering data about consciousness in others (under certain assumptions, e.g. reliability of verbal report), and relating it to the third-person data. And third, by careful analysis and thought-experiments etc, which can deliver useful conclusions. This way we at least get a stock of regularities and correlations at the "macroscopic" level to start with. then we try to boil these down to a simplest form and ultimately find "fundamental" principles that support these macroscopic regularities. A form of inference to the best explanation, if you like. That might be difficult and speculative, but not obviously impossible. It will be hard enough coming up with one good set of fundamental laws that fits the data, so if we do find a good hypothesis that explains the data we have, we'd have some good reason to believe it.

The panprotopsyichist can avail herself of all of this. The only extra thing is that the panprotopsyichist has strong constraints on the form of the fundamental laws. The basic psychophysical will ultimately connect fundamental physical dispositions to fundamental protophenomenal properties, right down at the microphysical level. Presumably there will be just a few such laws, one for every fundamental physical property.

That puts a lot of constraints on the form of the panprotopsyichist's fundamental psychophysical laws, which might either be a blessing or a curse. A blessing because the extra constraint narrows down the hypothesis space and tells us something about where to look. A curse because it may be all the harder to come up with laws that do the job. It's not at all obvious that we can derive all the facts about the macrophenomenology we know and love as a consequence of a whole bunch of protophenomenal properties we know and love. That's the biggest problem for the view -- what I call the "combination" or "constitution" problem in the book and elsewhere (see e.g. pp. 305-8). How does all those zillions of protophenomenal properties add up to a unified, bounded, grainless, structured phenomenal consciousness? It's not obvious that they can, but it's not obvious that they can't.

If this problem can be solved, then I think the panprotopsyichist view is the best solution to the mind-body problem by a mile. If it can't, then we have to look at other options. I'm hoping that we will get a better sense of the possible solutions to this problem for the panprotopsyichist in coming years.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 12 01:20:28 1999
Date: Fri, 12 Mar 1999 02:19:24 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Re: This and that
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I have a brief question regarding the admissability of causal primary intensions. This has come up at least twice; once in Tim's response to Josh's BIV scenario, and then again in our first class meeting. So, the problem is, I don't think primary intensions can ever be appropriately thought of as "the cause of my sensation that" or "that which brought about my experience of" or some such, because in any of these cases the

primary intension can vary wildly, depending on what the cause happens to be. So some brain in a vat may have a PI of water that appeals to electrodes and stimulation or whatever, and on earth this of course won't make any sense. So "that which causes my x experience" makes a primary intension the same as whatever the cause happens to be, and not the essential notion or properties of the thing itself, apriori. So I think the primary intension has to be necessarily connected to the essential quality of the thing (at least as it is known apriori or conceptually). What we mean by the concept of water, then, is not what causes the water experience (which may have nothing to do with the essence of water itself).

Anyway, this isn't a major point, but it keeps bothering me. I would be interested to hear any defence of causally-based primary intension, or where I have gone wrong or oversimplified the issues here.

Erik L.

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Mon Mar 1 06:51:23 1999
Date: Mon, 1 Mar 1999 06:51:13 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: next readings
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

The visa has been approved (hooray), but the paperwork still needs to come through, so I probably won't be there until next week. Meantime we have fallen behind a little and some people still haven't posted their week 5 contributions, so we can spend a little more time on that. I am particularly interested in comments on the general form of the argument as e.g. laid out in "Mind and Modality" -- if the argument is to be questioned, where will one question it? Which premise, or which reasoning?

Meantime we should move onto week 6 (the original week 7) readings sometime soon. This will concentrate around the PPR symposium on my book, in which all the respondents were type-B materialists, it turns out. Some of the material is on the web (all accessible via "online papers on consciousness"). I suggest starting with Yablo's "Concepts and Consciousness" (which is excellent); you could also take a look at his "Textbook Kripkeanism and the Open Texture of Concepts" which covers the same ground in a bit more detail, although this isn't as vital. You should also look at Brian Loar's important paper "Phenomenal States", not from the symposium, but which sets out a sophisticated type-B view which he reiterates in the symposium. Finally there is my response "Materialism and the Metaphysics of Modality" which goes over a lot of these issues more carefully.

The other three papers (Hill/McLaughlin, Loar, Shoemaker) are not online, and I will have to fax or express mail them over tomorrow. In the meantime the papers I mentioned should be a good starting point. Any comment on those papers and on the general dialectic will be welcome. Time for bed now.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Mar 1 21:22:16 1999
Date: Mon, 1 Mar 1999 21:20:45 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: readings
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Dear all,

I have put sent some papers by express mail to Ann Hickman. They should be there by Wednesday or Thursday. The papers include

- (1) PPR symposium papers by Hill & McLaughlin, Loar, Shoemaker, Yablo.
- (2) Chalmers, "Materialism and the Metaphysics of Modality".
- (3) Hill, "Imaginability, Conceivability, Possibility, and the Mind-Body Problem"
- (4) Levine, Review of _The Conscious Mind_.
- (5) Evans, Reference and Contingency
- (6) Jackson, Finding the Mind in the Natural World.
- (7) van Cleve, Conceivability and the Cartesian argument for dualism

(8) Yablo, Does Conceivability Imply Possibility?

The most important papers to look at immediately are the papers in (1). Yablo's paper in (1) is on the web, but I included it for completeness. The same goes for (2). (3) is a useful paper to look at in association with Hill & McLaughlin's paper in (1), as it elaborates one of the central ideas. (4) gives another type-B materialist reply to the sort of arguments I give. All of these are relevant to our current and forthcoming discussion of the modal argument and the dialectic with the type-B materialist.

(5) and (6) are papers relevant to previous issues. (5) is an important early paper by Evans on some ideas in the 2-D framework. (6) gives a very nice sketch by Jackson of some core issues relating metaphysics, modality, and materialism, and also talks about the role of the 2-D framework here. I think some people might find this particularly clear and useful.

(7) and (8) will be relevant to the forthcoming discussion of conceivability and possibility (coming up fairly soon).

I will ask Ann to either (a) make copies for everyone, or (b) leave the papers somewhere where everyone can grab them and photocopy them for themselves. [Probably it will be (b), but I'll try for (a).]

In the meantime people should have read the Loar and Yablo papers and my PPR reply from the web, and might start discussing those issues shortly. (Both the Loar and Yablo papers on the web are particularly good, and nice statements of a type-B materialist position.)

Right now I'm still hoping I might arrive by next Tuesday in time for class, but we'll see.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Mar 4 01:28:05 1999
Date: Thu, 4 Mar 1999 01:27:30 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: arrival
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Dear all,

Amazingly enough, the visa paperwork has made it to Arizona, and now I only need to wait for the FedEx and get the visa itself in Sydney. It looks likely that I will be arriving in Tucson at 3:59pm next Tuesday. I was hoping to be there in time for class, but I guess not. Still, I would like to meet once or even twice next week if that is possible. Once to have a general meeting to go over topics so far, and once for a full to discuss some of the papers I recently sent, especially the material in the PPR symposium and related issues about the 2-D modal argument against materialism. That would get things nicely caught up and under control in time for spring break, and ready for real momentum in the second half of term.

Would having a meeting at 5pm on Tuesday work for everyone? And is there another time that might work, e.g. sometime on Thursday? Please e-mail me to let me know whether you could make these meetings and what time would work best.

On other topics, I'm looking forward to seeing everyone's week 6 contributions as soon as possible (and also this week's minutes, of course). Looking at my web site log, I haven't seen as many hits on the relevant material from arizona.edu as I would have liked. Everyone should have downloaded and printed

<http://ling.ucsc.edu/~chalmers/modality.html>

and

<http://ling.ucsc.edu/~chalmers/princeton.html>

as both are very important. The second outlines a lot of relevant material for the course in a definitive-ish (although brief) form, and the former should clarify quite a few things more discursively. The former in particular is vital to this week's readings (especially section 3). Of course you should also be looking at the other web material (e.g. Loar and Yablo).

The papers I sent should be there very soon, and Ann will either copy them for everyone or leave them somewhere where you can grab them to copy.

I'm looking forward to finally seeing everyone soon!

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 12 13:19:02 1999
Date: Fri, 12 Mar 1999 14:10:53 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: zombie
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

This might be a silly question, and I apologize if we've already covered this:

In reading your response to Shoemaker (Section 4) I was struck by how much phenomenal zombies are like us (or we like them...). Am I right in thinking that zombies think that they *know what it's like* to see red, and might even run arguments for dualism on the strength of these beliefs?

You suggest that the judgments of zombies are a) unjustified and b) negatively rational. I'm not quite sure what you mean by 'negatively rational'-- is it just that they don't reason in a way we consider justified, but nevertheless manage to avoid errors? (If the results of their reasoning processes are the same as ours, might we not say that their beliefs are formed by a process that is at least as reliable as our own?)

I suppose this is basically the line that eliminativists take. Are we to be assured of the fact that we have phenomenal experiences on the basis of the intuitive arguments in Ch.1? If zombies have the same beliefs, isn't hard to resist a skeptic/eliminativist who claims that we are really no different?

Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 14 16:11:49 1999
Date: Sun, 14 Mar 1999 16:11:01 -0800

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Ideal apriority
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Angela writes:

>Also, regarding the concept of a priori, or rather the definition that 'x
>is knowable a priori if x can be known prior to experience,' I don't see
>why problems like the one presented by Godel's theorem aren't more
>troubling...if we have to idealize away from our own cognitive limitations
>to come up with an appropriate idea of 'in principle a priori
>knowability' such that we are conceiving of a being that is not subject
>to the sorts of mathematical counterexamples like Godel's theorem, then it
>seems as if our concept of a priori knowable breaks down...i.e., how are
>we to conceive of the distinction between what is knowable prior to
>experience and what is not for an idealized being like the one needed to
>avoid the Godel counterexample? It's not clear to me what such a being
>could or couldn't know without actually having to look...if a priori were
>defined strictly in terms of the analyticity of statements, then there
>seems to be less of a problem keeping the distinction between a priori and
>a posteriori straight when idealizing so far from our own case...but this
>doesn't seem to be the way Dave wants, centrally, to define the a
>priori...so is it the case that "knowable prior to experience" is more
>akin, or closer in kind to an idea like "analyticity" than I think, or
>what?

A minor point: I wouldn't say "knowable prior to experience here", but rather "knowable independent of experience". It may be that it takes experience to acquire the relevant concepts, for example, so some a priori P can only be known by experiencing beings. Strictly speaking, the crucial thing is that experience doesn't contribute to the *justification* of the knowledge in question, given that one has the concepts. One might better define things by saying that P is a priori when P can be known in a way such that experience does not contribute to the justification of P, or such that P has an experience-independent justification. (An example: perhaps one might hold that only a being with red experiences can have the concept of "red", in which case "all red things are colored" would only be knowable by experiencing beings, but might still be a priori.)

Anyway, your point is interesting. It's certainly true that if we idealize the notion of apriority, then it becomes harder to know whether a given P is a priori. I take it that we can agree that there is still a distinction between P such that a human can know it independently of experience, and P such that some being can know it independently of experience. Call the former "human apriority" and the latter "ideal apriority". Your point is that it is harder to know whether P is ideally a priori than whether it is humanly a priori.

Of course it may not be trivial to know whether a given P is even humanly a priori. Take for example some complex P in mathematics (or maybe philosophy), such that only a genius mathematician, in 200 years time, will come to know that P. Then we ordinary folk aren't in a good position to know whether P is a priori. But I guess one might say, at least some human can know whether P is a priori. At least, some human can know P in an a priori way, and from there it hopefully won't be too hard for them to conclude that P is knowable a priori (as long as they know that they know that P, and as long as they know that

they know it a priori). Whereas for ideal apriority, maybe none of us can know.

I think this is a consequence of the notion of ideal apriority that we just have to accept. Presumably there are some P such that no human can know P, but such that smarter beings could know P independently of experience (e.g., P = the Godel sentence of the human mathematical community, or some such). Then P is ideally a priori, but we can never know that. This is to say that ideal apriority can be "epistemically opaque", in that we can't always know whether P is ideally a priori.

It's not clear to what extent this is really an objection to the notion. Plenty of properties are epistemically opaque. E.g., truth is epistemically opaque -- we can't always know whether P is true. Similarly, necessity is epistemically opaque, and so on. It may still be that we can chart interesting relations between these notions, make argument that particular statements are or are not true/necessary/apriori, and so on.

One objection might be that the epistemic opaqueness makes the notion less useful, and makes in particular makes it harder to judge its application in a given case. E.g., one might object that one could never know, for a given P, that P is not a priori -- we might know that *we* don't know it a priori, but not that no being could. Still, this doesn't seem to be a problem for many P. E.g. let P be "Bill Clinton is the 42nd president of the USA". It seems pretty clear that no being could know this a priori. That is, we are justified in saying this is not ideally a priori, even though we can't simulate the thinking of the smart beings in question.

I think the moral is that although we can't know exactly what smarter beings could know, so we can't know exactly what is ideally a priori, one can still come to justified conclusions about many things that such beings could and couldn't know, so one can still come to justified conclusions about what is and isn't ideally a priori. For example, we can know that " $2+2 = 4$ " is ideally a priori, and that "there is a telephone on this desk" is not.

I suppose the central question in our context is whether "there are no zombies" (or strictly, " $P \rightarrow Q$ " where P is the complete physical truth and Q a phenomenal truth) is more like the case of the Godel sentence or the case of the telephone. If one accepts the epistemic intuitions, it certain does not *seem* to be a priori, and it seems that *we* can't know it a priori. But maybe it is ideally a priori, in that a less limited being could know it a priori? In which case our epistemic intuitions here merely reflect our cognitive limitations?

That's a deep question. It corresponds to the possibility of what I was calling on Tuesday "type-C materialism", the idea that there is an a highly nonobvious a priori entailment from physical to phenomenal (so that zombies are prima facie conceivable but not ideally conceivable, etc). I don't have anything definitive to say about this.

I do hold out hope that it is in some ways more like the telephone case than the Godel sentence case, though. That is, I think we have some reason to believe that it is not just not a priori for us, but it is not a priori for any being. There is the point, for example, that if there is even an ideal a priori entailment from physical to phenomenal here, then there must be some sort of a priori conceptual

analysis of phenomenal concepts in functional and structural terms. And I think we are in a position to say with justification that any such analysis would be a misanalysis -- it simply gets the meaning of the concept wrong. Of course an opponent could respond by saying that I only say this because of my cognitive limitations, and that if I was less limited I could see the deep a priori equivalence between phenomenal and functional concepts. The dialectic gets a little fuzzy around here, but my own view is that even if we can't rule out type-C materialism definitively, there is a strong prima facie case against it.

(Incidentally this brings out why doing things in terms of "analyticity" might not help here. If the type-C materialist were right, maybe there could be an analytic but highly nonobvious equivalence between phenomenal and functional concepts. We already know that analytic truths aren't always trivial -- witness the tricky analysis of the concept of "knowledge". So maybe it could turn out that analyticity is epistemically opaque, just as ideal a priority is.)

That being said, it's a complex issue and a very worthwhile avenue to explore. If anyone wants to follow up on this, e.g. writing a paper exploring the type-C materialist position and the slack opened up in the argument by the difference between prima facie and ideal conceivability, etc, I'd be delighted.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 14 16:26:06 1999
Date: Sun, 14 Mar 1999 16:25:56 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: zombie
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Anthony writes:

>In reading your response to Shoemaker (Section 4) I was struck by how much
>phenomenal zombies are like us (or we like them...). Am I right in
>thinking that zombies think that they *know what it's like* to see red,
>and might even run arguments for dualism on the strength of these beliefs?

That's right. I discuss that at length in Chapter 5 of the book (which we haven't covered). I call it the "Paradox of Phenomenal Judgment". It certainly leads to some interesting tensions, though I argue that it doesn't lead to any fatal flaws.

>You suggest that the judgments of zombies are a) unjustified and b)
>negatively rational. I'm not quite sure what you mean by 'negatively
>rational'-- is it just that they don't reason in a way we consider
>justified, but nevertheless manage to avoid errors? (If the results of
>their reasoning processes are the same as ours, might we not say that
>their beliefs are formed by a process that is at least as reliable as our
>own?)

By "negatively rational", I mean that they don't make mistaken inferences, errors of reasoning, etc. Nevertheless, they are still wrong, and unjustified, at least on my view, as they don't have the positive evidence that we do that justifies our belief in experience. (The evidence is provided by experience itself!) I'm not sure I'd put

things in terms of reliability, exactly, as one can argue that the zombie's belief-formation process is unreliable here -- after all, it produces a lot of false beliefs!

>I suppose this is basically the line that eliminativists take. Are we to
>be assured of the fact that we have phenomenal experiences on the basis of
>the intuitive arguments in Ch.1? If zombies have the same beliefs, isn't
>hard to resist a skeptic/eliminativist who claims that we are really no
>different?

Someone could certainly try to do that. In response I'll say two things. First, I don't think zombies have exactly the same beliefs as us here, as many of our phenomenal concepts and beliefs are themselves partly constituted by our experiences, experiences which the zombie lacks. (See "Mind and Modality" lecture 3 here.) But even setting that point aside, I think the skeptic/eliminativist would only have a good argument here under the assumption that the justification of a belief always inheres in other beliefs. If so, and if my zombie twin had the same beliefs, I'd be in trouble. But I think it is independently implausible that the justification of our beliefs always inheres in other beliefs. In particular, I think it is plausible that beliefs are often justified by **experiences**, and that experiences provide direct evidence for our beliefs.

If that's right, then my phenomenal beliefs have justification that the zombie lacks: the zombie is in a very different epistemic situation, and lacks all my evidence. If so, then the mere possibility of a zombie forming the same beliefs in an unjustified way doesn't undermine the justification of my own beliefs here. I still know I have experiences, because of my first-person evidence! Now maybe the zombie **thinks** he has that evidence too, but he is wrong and unjustified in thinking so. Again, it's the experiences themselves that justify all the beliefs in question.

Anyway, this is a subtle and interesting dialectic, though a bit outside the issues we are focusing on here. There's a lot more on it in Chapter 5 of the book and in "The Content and Epistemology of Phenomenal Belief" (online).

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 5 14:28:45 1999
Date: Fri, 5 Mar 1999 15:23:41 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Brad J Thompson <bradt@U.ARIZONA.EDU>
Subject: Comments on Loar and Hill/McLaughlin
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Here are some thoughts about the strategy taken by Loar and Hill/McLaughlin regarding the a posterioricity of psychophysical identities.

I agree with Chalmers that it is not enough for the type-B materialist to explain why zombies are conceivable. They also have to give an account of why the conceivability in such cases is an unreliable guide to possibility.

I don't think that the cognitive accounts that Loar and H & M give are **as** irrelevant to the issue of whether conceivability is a guide to possibility as Chalmers suggests. Chalmers contends that we could give a

psychological account of the conceivability of any state of affairs (such as mathematical beliefs). He claims further that Loar's notion of a recognitional concept does no extra work beyond establishing that physical and phenomenal concepts are cognitively distinct. But there does seem to be the difference, as Loar discusses, that phenomenal concepts pick out their referents directly. This is a feature unique to phenomenal concepts. Agreeing again that a cognitive explanation alone does not fully address the anti-materialist argument, I do think that the case of phenomenal concepts can be distinguished from all other concepts in a way that gives the materialist a foot-hold for developing a response.

But what more can the materialist say in order to motivate his or her position? Unless given strong reasons to the contrary, we ought to accept that conceivability entails possibility. Loar and H & M seem to accept this burden. (There is more to be said here about *who* has the burden of proof--the materialist or the anti-materialist. The above claim provides reason to think that the materialist has the burden. I suspect others will disagree--I'd be interested in any arguments for the contrary position.)

Loar's conclusion seems to be rather weak--for all we know, phenomenal concepts pick out the same properties as physical concepts. But how could we discover or determine that they do in fact have the same referents? It isn't clear to me what Loar has to say about this, especially give his agreement that identity statements involve a prior analyses.

H & M suggest that the materialist view has greater coherence, simplicity, and does not violate other intuitive doctrines such as the causal efficacy of the mental. Chalmers suggests that these points are being forced to do double-duty here. I have to go now (!) so I won't discuss this issue, though I hopefully I can pick up on it later since it seems crucial if we grant that the conceivability of zombies has an *interestingly* distinctive cognitive explanation.

Brad

Brad J Thompson
bradt@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 19 13:10:27 1999
Date: Fri, 19 Mar 1999 14:09:09 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Comments on Yablo
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some comments on Yablo,

(1) I have some worries about the reliability of conceivability that Yablo either doesn't mention, or doesn't take as seriously as he should. The first is the fact that one's modal intuitions can be influenced by how the scenario that one is being asked to imagine is described. The locus classicus of this is Bernard William's wonderful paper *The Self and the Future*. In the first scenario, information from my brain is extracted and placed in your brain, and vice versa. Most people have a very strong intuition that my identity goes with my information, and thus the continuity of one's body is not sufficient for one's continued existence. In the second scenario, I am told that I will be tortured tomorrow, but

not to worry, for I will be given a drug before hand such that all of my memories are erased. Most people have an equally strong intuition that I should worry, and thus they think that continuity of one's body is sufficient for one's continued existence. (I'm glossing over a lot of the details here, hopefully most will be familiar with the paper.)

The moral that at least some have drawn from this is that, with a few exceptions, thought-experiments (or thought-experiments that go beyond what we know to be nomologically possible) are unreliable. (Kathleen Wilkes says *something* like this, as does Mark Johnston.)

There are a number of things that Yablo might say about this case. His first line of response (p. 38) doesn't look promising - nobody is conflating metaphysical possibility or conceivability with epistemic possibility here. He might want to push the second line (p. 39), for he says that 'it is all too easy to believe that much of the current controversy over conditions of personal identity and survival. . . owes more to our meaning slightly different things by "person" and "survive" than to any real clash of modal intuitions" (p. 39) But this doesn't look like it will work here, for it is one and the same person who can conceive of a scenario that supports P, and can also conceive of a scenario that also supports not-P. Perhaps what he should say is something like the following: whether one's ability to conceive of P provides good reason for thinking that P is (metaphysically) possible depends on the situation/background against which one is imagining P (see p. 28). The difference between the two scenarios is the context in which the P fact is imagined. To use an analogy, the difference is something like that between imaging a tiger in a zoo and imaging a tiger in a wild, and finding that in the first situation one can imagine the tiger to be polka-dotted, but in the second situation one cannot. And this of course is perplexing, because whether or not tigers can be polka-dotted should not depend on their immediate environment. I take it that Yablo needs to argue that the imagined context in Williams's scenario's is not as benign as putting one's imagined tiger in the wild or in the zoo - he needs to argue that it actually changes what it is that one is imagining. Here's another example. it would be a bad result for perception if two lines drawn on a piece of paper looked parallel during the week, but appeared to curve on weekends. But it's not so bad when two lines that appear to be parallel when viewed by themselves, seem to curve when embedded in a larger picture of other lines that curve. Then we have grounds for explaining the latter case as one of perceptual illusion. Yablo needs to explain why the context of one (or both) of Williams's scenarios creates a modal illusion. Of course, he may well be able to do that, but no-one seems to have had much success at this task.

(2) A second worry that I have is one that I've bought up before. It's the problem of inter-subjective disagreement concerning essentialism about one's own identity. Yablo several times refers to the Kripkean intuition that one's origins are essential to one. Some (Aristotelians?) some to think that although one's origins are not essential to one, one's species membership is. Others (an undergrad professor of mine) think that neither one's origins nor one's species membership are essential to one. He thought that he could conceive of himself as being a poached egg (an ordinary poached egg, not an eggy creature that walks and talks). Assuming some form of modal factualism, on which at most one of these positions is correct, how do we decide which one is? Again, it seems to me that neither (1) nor (2) (see p. 39) will work here. Will (3) work? Are there defeaters for one or more of these positions? Perhaps this is where the argument would turn. Perhaps the Kripkean would claim that I couldn't have been a poached egg, for if I could, then you might just as easily have been the same poached egg, but, necessarily, I couldn't have been you. But what is

one's reason for thinking that the last claim is true? That one cannot imagine a situation in which I am you? But that seems to be just what is in question. Yablo could also argue that the disagreement between the Kripkeans, Aristotelians and the poached egg people is contaminated by theoretical commitments, but again, this would have to be argued.

Now, of course, there are lots of modal problems that (almost) everyone agrees on, e.g., I could have had something different for breakfast this morning. So the argument cannot be that people fail to exhibit agreement on a lot of modal statements. But to say this isn't to say much. The problem seems to be that there are lots of important modal statements that people don't agree on, and it's not clear that they are making some non-modal factual error or logical error that would explain their modal disagreement (p. 39). The situation here seems to be akin to that of religious discourse, and to some extent moral discourse. Most people agree on most moral judgments, but that doesn't make the job of trying to understand why there is a residue of basic and fundamental moral disagreement any easier.

(3). Why does Yablo think that 'it is inconceivable that addition facts should vary between possible worlds'? (p. 32) Presumably it's not because of some inductive generalization. It's not as if he's tried to conceive of $\langle 2+3=5 \rangle$ in one world, and $\langle 2+3 \text{ not}=5 \rangle$ in some other world and failed, tried to imagine $\langle 3+3=6 \rangle$ in one world and $\langle 3+3 \text{ not}=6 \rangle$ in some other world and failed, and so on, and thus induced that all arithmetical statements have their truth-values necessarily. Has he directly tried to imagine all arithmetical truths having their truth-values contingently? How would one do that? I have some sense of what is involved in entertaining the thought $\langle \text{all arithmetical statements have their truth-values contingently} \rangle$, but I have no idea of how to submit this thought to imaginative scrutiny.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 28 21:58:52 1999
Date: Sun, 28 Mar 1999 22:57:50 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Angela J Burnette <aburnett@U.ARIZONA.EDU>
Subject: Re: Loar's mode of presentation argument
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Loar's talk of different modes of presentation...phenomenal states are presented, or represented in our concepts differently from physical states, which are mediated by perception of extrinsic qualities rather than directly presented by intrinsic qualities...seems to be offering an explanation of why we can (and maybe necessarily do) have different concepts for possibly the same properties, but I'm not clear on exactly where Loar's mode of presentation considerations are supposed to push on the anti-materialist argument...

is he claiming that the conceivability of zombie worlds doesn't warrant the move from conceivability to possibility because we are fooled into

thinking that they can actually come apart just because we can conceive of them coming apart (due to the fact that they are presented via different modes?) In other words, is he offering evidence to the effect that conc. doesn't nec. track possibility given the unique quality of phenomenal states? because it seems to me that the ideal conceiver would not be mistaken by different modes of presentation, but as this has not come up as a response to Loar (to my knowledge) I wonder if I am understanding the argument or not....

angela

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 28 23:37:53 1999

Date: Mon, 29 Mar 1999 00:34:58 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik J Larson <erikl@U.ARIZONA.EDU>

Subject: Re: Loar's mode of presentation argument

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

I take it that Loar is making the point that the conceptual independence of the phenomenal and physical is consistent with psychological differences and doesn't necessarily tell us anything about differences in expressed properties, which is just to say that the epistemic premise concerning conceptual distinctness does not entail the metaphysical conclusion concerning anti-materialism. So I take it that Loar's "mode of presentation" argument is intended as a counterexample to the move from conceivability to possibility.

I accept Chalmer's reply to Loar, i.e., that he presupposes strong psychophysical necessities rather than explaining them. So I'm not so worried about Loar, specifically. My worry is more generally the move from an epistemic premise to a metaphysical conclusion. Loar points out one possible way in which we may fail to draw solid metaphysical conclusions from epistemic considerations. The argument may be flawed in its specific guise. But I think there is a general confusion here, about exactly what the anti-materialist argument can sustain metaphysically given the epistemic starting point. I am tempted to say that the (anti-materialist) argument establishes this: Because we cannot see any necessary connection between physical and phenomenal states--the conditional $P \rightarrow Q$ is not apriori--we cannot positively establish the truth of materialism. I take it that the argument is stronger: Because physical and phenomenal states necessarily have different primary intensions, materialism is false (add in "there is a logically possible world where P holds but not Q and so on). The second formulation of the argument gets to the conclusion, but it seems that it rests on much shakier epistemic ground.

Maybe we should add some antecedent clause, to the effect that "If we have access to the correct primary intensions of phenomenal and physical states, and if we are able to completely specify (without fear of confusion or cognitive obscurities etc) these intensions...", then we get necessarily distinct primary intensions and a positively conceivable zombie scenario and a good foothold on the anti-materialist conclusion. But for myself, all I can see is the former argument-- we cannot see any necessary connection between physical and phenomenal states. And this is just a fact about what we can and can't see, and not much more. So I'm looking for someone to pull me out of the epistemic trap, so to speak (I'm guessing it will be Dave).

Erik

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From bayne@U.Arizona.EDU Fri Mar 19 14:25:30 1999
Date: Fri, 19 Mar 1999 15:24:31 -0700 (MST)
From: Timothy J Bayne <bayne@U.Arizona.EDU>
To: David Chalmers <chalmers@ling.ucsc.edu>, MODALITY@LISTSERV.ARIZONA.EDU
Subject: A Comment on Hill
Status: RO

Hill makes some interesting points about having access to the same property via two different modalities. I take it that he doesn't want his argument to rest on the distinction between a property and its mode of presentation (65). Rather, he wants to exploit the distinction between two types of imagination. So I **think** that he wants to claim that a property can be presented to types of imagination under the same mode of presentation, and yet these two presentations differ in their intrinsic character. I'm not sure his case for this claim (if it is his claim) can be maintained.

As I read him, Hill attempts to motivate this line by arguing that the same property (under the same mode of presentation?) can be presented to two different senses, and yet these two presentations can differ in intrinsic character: "Is it possible for psychological mechanisms to produce radically different experiences if those experiences are in fact presentations of the same property? Well, yes, this happens all of the time in the case of perceptual presentations. Compare a visual presentation of the surface of a piece of sandpaper with a tactual presentation of the same surface. Any two such presentations will be quite different in point of intrinsic character; but still the properties that are presented by the former will overlap with the properties that are presented by the latter (p. 67)." The trouble with this example is that in seeing a piece of sandpaper one is given a lot of information about it that one does not get in touching it, and vice-versa, and arguably that is what makes the experience of seeing that something is shaped like X different from feeling that it is shaped like X.

What we want is a case in which one **just** gets information about the paper's surface structure, and the nature of this experience is different depending on the modality, without the mode of presentation being different. It is not obvious that in such a case there would be a difference between the visual-experience and the tactile experience. (Part of the problem here is knowing what it would be like to **just** see the surface structure, or just feel the surface structure - but leave that aside.) Despite his avowed intentions, Hill seems to be simply back into the distinction between a property and its mode of presentation: there's the one property of the surface structure of the paper, and it has two modes of presentation. Seeing that something is square, and feeling that it is square may both involve the single property of squareness, but it also seems to involve two distinct modes of presentation: feeling-square and looking-square. 'Feeling-square' and 'looking-square' seem to be

distinct properties.

This issue is basically Molyneaux's question to Locke: suppose that a person blind from birth were suddenly to regain their sight, would they be able to identify objects without touching them? Or would they have to discover that feeling-square and looking-square both converged on a single property, being square? Put the issue another way: is it contingent that something looks square iff it feels square? (Perhaps we should relativise this to ideal circumstances: X looks square under ideal conditions iff it feels square under ideal conditions.

It seems important to distinguish between two types of sensory concepts here. The first type refers to properties that are accessible to different modalities, but are such that the different modalities experience them in such a way that the perceiver may not realize (even under ideal circumstances) that she is experiencing the same property. Surface structure, or squareness, may be such a property. The second type refer to properties that seem to be accessible, in some sense, to only one sense modality. (Phenomenal) red and the sounds of middle C might be such properties. Although the concept <red> and <middle C> are not "tied" or "restricted" to particular modalities - arguably they wouldn't be *concepts* if their inferential range were restricted to a modality - they do seem to be grounded in a single modality in the way that, say, <square> is not (even for a blind person).

But one might come back and say that the property red can be directly detected on the basis of non-visual cues - this is what Mary can do. The counter-reply is that Mary can use some kind of red concept, but not her concept of phenomenal red, which refers to the qualia. But now one wants to know about <visual squareness> and <tactile squareness>. Exactly how are these related? they seem to be sensory concepts, but they don't seem to be modality specific in the way that, say, <phenomenal red> is. What one attempts to refer to in thinking that X looks square is a fact about X that is also directly accessible to other modalities. Arguably, this is not the case with (phenomenal) red. In judging that X looks red, there is (arguably!) no commitment to thinking that that same property can be directly identified via some other mode of access. I take it that Hill would not like this result. He would want to resist the idea of a property (phenomenal-red, visual-squareness) that is logically modality specific.

The upshot: Hill's "2-types of imaginability" position seems to collapse into, or at least require, a "two modes of presentation" position - and he doesn't want that (I take it).

cheers,

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 14 16:57:22 1999
Date: Sun, 14 Mar 1999 16:57:11 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Readings, etc
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Someone asked that I post the readings for next meeting (after spring break). This meeting will be "week 8" from the syllabus, on "Does Conceivability Entail Possibility?". The main readings are:

Chalmers, Mind and Modality, part 2.
Yablo, Is conceivability a guide to possibility?
van Cleve, Conceivability and the Cartesian argument for dualism

We won't be going over the Yablo and van Cleve in huge detail, but they provide very interesting and useful background material for the general issues here, so you should go over them carefully. In seminar I will be developing the various distinctions between different kinds of conceivability in section 2.3 of the Princeton lectures, and will probably also go over the varieties over modal rationalism in 2.4. So you should concentrate especially on those sections and try to understand them and come with any questions. We'll be going over the material in 2.5 and later in coming weeks.

You should also read the Hill paper on conceivability from Philosophical Studies, if you haven't already, and Loar's paper "Phenomenal States", as these are both quite relevant.

We've talked a lot on the mailing list about general issues about consciousness and the 2-D framework, but I'd like to see us start to concentrate a bit more on the issues re modality, modal arguments, the epistemic/modal bridges, strong necessities, modal space and modal rationalism, and so on (i.e., the issues in "week 6" onward). I'd particularly like to see people getting into the details. It's good to have a broad understanding, but at the end of the day the details are where the cash value is, and there are all sorts of interesting things going on in and around the details here. And I'll be expecting term papers that really go into things carefully. So in the meantime I'd be pleased to see comments on the ins and outs of the arguments, distinctions, and the general dialectic here. That goes both for the material I've presented and for the material in the readings. The mailing list might be a particularly good place to go into these things, as it's arguably easier to go into the details carefully outside real time.

It might be good to meet twice again the week after spring break, if we can (week 9 is very light on reading, so it would make sense to schedule a double-whammy then). I go out of town on Thursday night, but Thursday at 3 would probably work for me. Let me know if it doesn't work for any of you.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 19 13:10:27 1999
Date: Fri, 19 Mar 1999 14:09:09 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Comments on Yablo

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Some comments on Yablo,

(1) I have some worries about the reliability of conceivability that Yablo either doesn't mention, or doesn't take as seriously as he should. The first is the fact that one's modal intuitions can be influenced by how the scenario that one is being asked to imagine is described. The locus classicus of this is Bernard Williams's wonderful paper **The Self and the Future**. In the first scenario, information from my brain is extracted and placed in your brain, and vice versa. Most people have a very strong intuition that my identity goes with my information, and thus the continuity of one's body is not sufficient for one's continued existence. In the second scenario, I am told that I will be tortured tomorrow, but not to worry, for I will be given a drug before hand such that all of my memories are erased. Most people have an equally strong intuition that I should worry, and thus they think that continuity of one's body is sufficient for one's continued existence. (I'm glossing over a lot of the details here, hopefully most will be familiar with the paper.)

The moral that at least some have drawn from this is that, with a few exceptions, thought-experiments (or thought-experiments that go beyond what we know to be nomologically possible) are unreliable. (Kathleen Wilkes says **something** like this, as does Mark Johnston.)

There are a number of things that Yablo might say about this case. His first line of response (p. 38) doesn't look promising - nobody is conflating metaphysical possibility or conceivability with epistemic possibility here. He might want to push the second line (p. 39), for he says that 'it is all too easy to believe that much of the current controversy over conditions of personal identity and survival. . . owes more to our meaning slightly different things by "person" and survive' than to any real clash of modal intuitions" (p. 39) But this doesn't look like it will work here, for it is one and the same person who can conceive of a scenario that supports P, and can also conceive of a scenario that also supports not-P. Perhaps what he should say is something like the following: whether one's ability to conceive of P provides good reason for thinking that P is (metaphysically) possible depends on the situation/background against which one is imagining P (see p. 28). The difference between the two scenarios is the context in which the P fact is imagined. To use an analogy, the difference is something like that between imaging a tiger in a zoo and imaging a tiger in a wild, and finding that in the first situation one can imagine the tiger to be polka-dotted, but in the second situation one cannot. And this of course is perplexing, because whether or not tigers can be polka-dotted should not depend on their immediate environment. I take it that Yablo needs to argue that the imagined context in Williams's scenario's is not as benign as putting one's imagined tiger in the wild or in the zoo - he needs to argue that it actually changes what it is that one is imagining. Here's another example. it would be a bad result for perception if two lines drawn on a piece of paper looked parallel during the week, but appeared to curve on weekends. But it's not so bad when two lines that appear to be parallel when viewed by themselves, seem to curve when embedded in a larger picture of other lines that curve. Then we have grounds for explaining the latter case as one of perceptual illusion. Yablo needs to explain why the context of one (or both) of Williams's scenarios creates a modal illusion. Of course, he may well be able to do that, but no-one seems to have had much success at this task.

(2) A second worry that I have is one that I've bought up before. It's the

problem of inter-subjective disagreement concerning essentialism about one's own identity. Yablo several times refers to the Kripkean intuition that one's origins are essential to one. Some (Aristotelians?) some to think that although one's origins are not essential to one, one's species membership is. Others (an undergrad professor of mine) think that neither one's origins nor one's species membership are essential to one. He thought that he could conceive of himself as being a poached egg (an ordinary poached egg, not an eggy creature that walks and talks). Assuming some form of modal factualism, on which at most one of these positions is correct, how do we decide which one is? Again, it seems to me that neither (1) nor (2) (see p. 39) will work here. Will (3) work? Are there defeaters for one or more of these positions? Perhaps this is where the argument would turn. Perhaps the Kripkean would claim that I couldn't have been a poached egg, for if I could, then you might just as easily have been the same poached egg, but, necessarily, I couldn't have been you. But what is one's reason for thinking that the last claim is true? That one cannot imagine a situation in which I am you? But that seems to be just what is in question. Yablo could also argue that the disagreement between the Kripkeans, Aristotelians and the poached egg people is contaminated by theoretical commitments, but again, this would have to be argued.

Now, of course, there are lots of modal problems that (almost) everyone agrees on, e.g., I could have had something different for breakfast this morning. So the argument cannot be that people fail to exhibit agreement on a lot of modal statements. But to say this isn't to say much. The problem seems to be that there are lots of important modal statements that people don't agree on, and it's not clear that they are making some non-modal factual error or logical error that would explain their modal disagreement (p. 39). The situation here seems to be akin to that of religious discourse, and to some extent moral discourse. Most people agree on most moral judgments, but that doesn't make the job of trying to understand why there is a residue of basic and fundamental moral disagreement any easier.

(3). Why does Yablo think that 'it is inconceivable that addition facts should vary between possible worlds'? (p. 32) Presumably it's not because of some inductive generalization. It's not as if he's tried to conceive of $\langle 2+3=5 \rangle$ in one world, and $\langle 2+3 \text{ not}=5 \rangle$ in some other world and failed, tried to imagine $\langle 3+3=6 \rangle$ in one world and $\langle 3+3 \text{ not}=6 \rangle$ in some other world and failed, and so on, and thus induced that all arithmetical statements have their truth-values necessarily. Has he directly tried to imagine all arithmetical truths having their truth-values contingently? How would one do that? I have some sense of what is involved in entertaining the thought $\langle \text{all arithmetical statements have their truth-values contingently} \rangle$, but I have no idea of how to submit this thought to imaginative scrutiny.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 23 22:42:43 1999
Date: Tue, 23 Mar 1999 22:41:40 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Admin
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I e-mailed Alvin about using the library tomorrow. He was planning on using it 2-4, and suggested we split the difference and meet at 3:45. So that can be our tentative plan. It occurred to me also that we didn't consider the possibility of meeting earlier, e.g. at 2 or even at 12 (that way maybe Thony could make it). It probably won't be practical to make such a change even if theoretically possible, but just in case, could everyone e-mail me immediately after getting this and tell me whether an earlier time would work?

I'll be interested to hear everyone's thought on the issues about conceivability and possibility over e-mail. There are lots of interesting open issues here to discuss -- e.g. concerning (1) the definition of different kinds of conceivability (ideal conceivability, positive conceivability, etc); (2) the gap between prima facie and ideal conceivability, and our epistemic access to the latter; (3) the issue of how good a guide secunda facie conceivability is to possibility, and whether there are some interesting counterexamples (I'm very interested to hear of any!); (4) the difference between negative and positive conceivability and whether there is anything in the "twilight zone" between these (i.e., ideally negatively conceivable but not ideally positively conceivable); and so on. All thoughts on these and other relevant issues are welcome!

I should remind people that this is just like any other class where writing is due every week; not handing in work is a black mark. Now that we have real meetings, I'll relax the requirement that everyone should post at least one non-reply contribution per week. At least one reasonably substantial contribution is expected per week, but it can be a reply to or an extension of someone else's contribution.

The only reading for tomorrow are sections 2.5 and 2.6 of Mind and Modality (which hopefully you've read already). For next week, the readings are

TCM, Section 2.5

Block & Stalnaker, Conceptual analysis and the explanatory gap [web]

Byrne, Cosmic hermeneutics.

Horgan, Supervenience and cosmic hermeneutics

Block and Stalnaker is available on the web via my "online papers on consciousness" page. I have put Byrne and Horgan in the purple folder in the department office; make yourself a copy.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 19 13:10:27 1999
Date: Fri, 19 Mar 1999 14:09:09 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Comments on Yablo
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some comments on Yablo,

(1) I have some worries about the reliability of conceivability that Yablo

either doesn't mention, or doesn't take as seriously as he should. The first is the fact that one's modal intuitions can be influenced by how the scenario that one is being asked to imagine is described. The locus classicus of this is Bernard Williams's wonderful paper *"The Self and the Future"*. In the first scenario, information from my brain is extracted and placed in your brain, and vice versa. Most people have a very strong intuition that my identity goes with my information, and thus the continuity of one's body is not sufficient for one's continued existence. In the second scenario, I am told that I will be tortured tomorrow, but not to worry, for I will be given a drug before hand such that all of my memories are erased. Most people have an equally strong intuition that I should worry, and thus they think that continuity of one's body is sufficient for one's continued existence. (I'm glossing over a lot of the details here, hopefully most will be familiar with the paper.)

The moral that at least some have drawn from this is that, with a few exceptions, thought-experiments (or thought-experiments that go beyond what we know to be nomologically possible) are unreliable. (Kathleen Wilkes says *"something"* like this, as does Mark Johnston.)

There are a number of things that Yablo might say about this case. His first line of response (p. 38) doesn't look promising - nobody is conflating metaphysical possibility or conceivability with epistemic possibility here. He might want to push the second line (p. 39), for he says that 'it is all too easy to believe that much of the current controversy over conditions of personal identity and survival. . . owes more to our meaning slightly different things by "person" and "survive" than to any real clash of modal intuitions" (p. 39) But this doesn't look like it will work here, for it is one and the same person who can conceive of a scenario that supports P, and can also conceive of a scenario that also supports not-P. Perhaps what he should say is something like the following: whether one's ability to conceive of P provides good reason for thinking that P is (metaphysically) possible depends on the situation/background against which one is imagining P (see p. 28). The difference between the two scenarios is the context in which the P fact is imagined. To use an analogy, the difference is something like that between imagining a tiger in a zoo and imagining a tiger in the wild, and finding that in the first situation one can imagine the tiger to be polka-dotted, but in the second situation one cannot. And this of course is perplexing, because whether or not tigers can be polka-dotted should not depend on their immediate environment. I take it that Yablo needs to argue that the imagined context in Williams's scenario's is not as benign as putting one's imagined tiger in the wild or in the zoo - he needs to argue that it actually changes what it is that one is imagining. Here's another example. it would be a bad result for perception if two lines drawn on a piece of paper looked parallel during the week, but appeared to curve on weekends. But it's not so bad when two lines that appear to be parallel when viewed by themselves, seem to curve when embedded in a larger picture of other lines that curve. Then we have grounds for explaining the latter case as one of perceptual illusion. Yablo needs to explain why the context of one (or both) of Williams's scenarios creates a modal illusion. Of course, he may well be able to do that, but no-one seems to have had much success at this task.

(2) A second worry that I have is one that I've brought up before. It's the problem of inter-subjective disagreement concerning essentialism about one's own identity. Yablo several times refers to the Kripkean intuition that one's origins are essential to one. Some (Aristotelians?) seem to think that although one's origins are not essential to one, one's species membership is. Others (an undergrad professor of mine) think that neither one's origins nor one's species membership are essential to one. He

thought that he could conceive of himself as being a poached egg (an ordinary poached egg, not an eggy creature that walks and talks). Assuming some form of modal factualism, on which at most one of these positions is correct, how do we decide which one is? Again, it seems to me that neither (1) nor (2) (see p. 39) will work here. Will (3) work? Are there defeaters for one or more of these positions? Perhaps this is where the argument would turn. Perhaps the Kripkean would claim that I couldn't have been a poached egg, for if I could, then you might just as easily have been the same poached egg, but, necessarily, I couldn't have been you. But what is one's reason for thinking that the last claim is true? That one cannot imagine a situation in which I am you? But that seems to be just what is in question. Yablo could also argue that the disagreement between the Kripkeans, Aristotelians and the poached egg people is contaminated by theoretical commitments, but again, this would have to be argued.

Now, of course, there are lots of modal problems that (almost) everyone agrees on, e.g., I could have had something different for breakfast this morning. So the argument cannot be that people fail to exhibit agreement on a lot of modal statements. But to say this isn't to say much. The problem seems to be that there are lots of important modal statements that people don't agree on, and it's not clear that they are making some non-modal factual error or logical error that would explain their modal disagreement (p. 39). The situation here seems to be akin to that of religious discourse, and to some extent moral discourse. Most people agree on most moral judgments, but that doesn't make the job of trying to understand why there is a residue of basic and fundamental moral disagreement any easier.

(3). Why does Yablo think that 'it is inconceivable that addition facts should vary between possible worlds'? (p. 32) Presumably it's not because of some inductive generalization. It's not as if he's tried to conceive of $\langle 2+3=5 \rangle$ in one world, and $\langle 2+3 \text{ not}=5 \rangle$ in some other world and failed, tried to imagine $\langle 3+3=6 \rangle$ in one world and $\langle 3+3 \text{ not}=6 \rangle$ in some other world and failed, and so on, and thus induced that all arithmetical statements have their truth-values necessarily. Has he directly tried to imagine all arithmetical truths having their truth-values contingently? How would one do that? I have some sense of what is involved in entertaining the thought $\langle \text{all arithmetical statements have their truth-values contingently} \rangle$, but I have no idea of how to submit this thought to imaginative scrutiny.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 23 23:32:48 1999
Date: Tue, 23 Mar 1999 23:32:36 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Comments on Yablo
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some brief thoughts re Tim's thoughts on Yablo. All other comments

are welcome.

>(1) I have some worries about the reliability of conceivability that Yablo
>either doesn't mention, or doesn't take as seriously as he should. The
>first is the fact that one's modal intuitions can be influenced by how the
>scenario that one is being asked to imagine is described. The locus
>classicus of this is Bernard Williams's wonderful paper *The Self and the
>Future*. In the first scenario, information from my brain is extracted and
>placed in your brain, and vice versa. Most people have a very strong
>intuition that my identity goes with my information, and thus the
>continuity of one's body is not sufficient for one's continued existence.
>In the second scenario, I am told that I will be tortured tomorrow, but
>not to worry, for I will be given a drug before hand such that all of my
>memories are erased. Most people have an equally strong intuition that I
>should worry, and thus they think that continuity of one's body is
>sufficient for one's continued existence. (I'm glossing over a lot of the
>details here, hopefully most will be familiar with the paper.)

This is interesting. Of course not everyone has both intuitions. I'd
be surprised if "most people" have the first intuition, for example.
I think a lot of people intuitively think that identity goes with the
brain, so in the first case I won't survive in the new brain. And the
strength of the second intuition is unclear. But I suppose that *if*
one has both intuitions, then one seems to be suggesting that
continuity of info is sufficient for survival, and that continuity of
brain/body is sufficient for survival. As it stands, those intuitions
aren't quite contradictory -- maybe one can survive either way?
Though it would seem to commit one to the possibility of double
survival in certain cases, which seems odd (unless one has a rule for
deciding when both criteria are present in separate cases).

But anyway, let's ignore the non-contradictoriness and think about
what to say if we really did have two contradictory intuitions here:
i.e. we find A positively conceivable, B positively conceivable, where
the possibility of A contradicts the possibility of B. I think in
such cases, one can argue that one is being inconsistent in one's
evaluation of the conceivable scenarios, and that further rational
reflection will reveal this, so both won't turn out to be ideally
positively conceivable. If the A is truly (i.e. ideally) positively
conceivable, then analysis of this will reveal that B is not, and vice
versa. And indeed I think this mirrors our reaction to the cases.
Even if one has a momentary feeling that both intuitions are
reasonable, a bit of reflection reveals the conflict and leaves us at
best uncertain, or perhaps withdrawing one of the claims completely.

Of course that's just the case where both intuitions are had in one
person. Where conflicting intuitions re personal identity are had by
two different people, it's tricky. Just say A says he can conceive of
teletransportation being survival and B says he conceives of it being
death. I suppose that's not automatically a contradiction, as maybe
they are conceiving different "further facts" in addition to the micro
facts: if one is a nonreductionist about PI one might well think that
both survival and death are positively conceivable here.

So let's take a strong case, where A says teletransportation survival
is not conceivable, and B says that it is; and B says that
teletransportation death is not conceivable, where A says that it is.
These intuitions are clearly in conflict. One possibility is that A
and B have different concepts of "survival" and "death": perhaps B has
a deflationary concept of "functional survival", whereas A has an
inflationary concept of "intrinsic survival", for example. In that

case, we might find that on disambiguation, the two agree on the relevant intuitions (teletransportation is always functional survival, never intrinsic survival). Though it may be that B will say that intrinsic survival does not exist; in that case A and B will differ not on what's conceivable but on what's actual.

Another possibility is that both are to some degree irrational, or at least are making claims that are infected by theory or wishful thinking as much as direct conceivability. For my part, for example, I think that both teleport survival and teleport death are *prima facie* negatively conceivable, and I don't have strong intuitions about how best to describe the positively conceivable teleport scenario. So maybe A and B should best be agnostic, though they may have reasonably strong intuitions here. Or maybe it's a place where continued rational reflection can reveal that one is right and the other wrong.

Tim also raises the case where P is conceivable against one background but not against the other. In such a case I'm tempted to say that what's here at issue is *prima facie* conceivability, and that even *secunda facie* reflection indicates that one ought to treat the two cases equally. Maybe background can bias one in a certain direction *prima facie*, but this biasing role should be factored out on ideal reflection.

Tim's case of conflict over essentialism is tricky. As noted in class, this is a case of secondary conceivability, i.e. intuitions about what is secondarily possible (conceivable/possible in worlds considered as counterfactual). That's often a *posteriori*, but arguably should be a *a priori* conditional on the empirical non-modal facts being all in or not at issue. So what's the *a priori* disagreement here? Different concepts of person or object? Maybe, but doesn't seem too promising. Insufficient rational reflection? I think Kripke would say your teacher is just wrong, and not being properly rational, but of course it's not clear how to adjudicate that. Vagueness/indeterminacy/ambiguity? I'm tempted to say that the whole question of what is essential to X is somewhat vague, and I can often go that way. So maybe the fact of the matter is indeterminate, and the two are "precisifying" in different ways.

I'd like to say that insofar as the matter is determinate, it's settleable by rational reflection, and vice versa. So any irresolvable rational dispute (given that the empirical facts are in) corresponds to an indeterminacy. That's more or less the scrutability claim we'll be discussing, and it's somewhat controversial, but not wholly implausible, I think.

This also touches on the question of disagreement over *a priori* truths by apparently rational being. E.g., Graham Priest, a very smart philosopher, thinks contradictions are sometimes true. Most others think it is *a priori* (and obvious) that they never are. Is he being irrational? Well, he at least seems sane and a good reasoner. But I guess one has to say that at some level he is not being rational, even though he is to some extent "procedurally rational". Maybe some conceivability disagreements are analogous to this, where someone says something that is *a priori* false despite seeming procedurally rational. Maybe even the poached egg case, at a deep level?

So there are various resources for someone who endorses a conceivability-possibility link, but clearly none of this is obvious or trivial. All thoughts are welcome.

--Dave.

P.S. Re mathematics, I think Yablo thinks it is a priori that $3+3=6$, so it isn't conceivable otherwise. If P is a priori, not-P is not negatively conceivable or positively conceivable. In any case I take it that there's no positive reason to think it conceivable that $3+3$ is not 6; and the fact that it is a priori suggests that any scenario will verify " $3+3=6$ ". (Given that arbitrary a priori reasoning is allowed in seeing what verifies what.)

From owner-modality@LISTSERV.ARIZONA.EDU Wed Mar 24 07:01:31 1999
Date: Wed, 24 Mar 1999 07:59:10 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Re: Comments on Yablo
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R0

>
> P.S. Re mathematics, I think Yablo thinks it is a priori that $3+3=6$,
> so it isn't conceivable otherwise. If P is a priori, not-P is not
> negatively conceivable or positively conceivable. In any case I take
> it that there's no positive reason to think it conceivable that $3+3$ is
> not 6; and the fact that it is a priori suggests that any scenario
> will verify " $3+3=6$ ". (Given that arbitrary a priori reasoning is
> allowed in seeing what verifies what.)

This seems to me to get things around the wrong way. We think that $3+3=6$ is a priori *because* its negation is inconceivable, not vice-versa. This is a bit clearer with respect to geometrical truths, it seems to me. Why did Kant think that Euclidian geometry was a priori true of this world? (It's not as if it comes with a label stuck to it: 'a priori'.) He tried to conceive (imagine? visualize?) a physical world in which geometry was non-Euclidean, and failed. It seems to me that thinking P is a priori involves attempting to conceive of not-P, and being unable to. (Perhaps that should be: one is justified in thinking that p is a priori if (only if?) one has tried to conceive of not-P and been unable to.

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Mar 24 09:40:05 1999
Date: Wed, 24 Mar 1999 10:39:06 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Mary and Negative possibility
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R0

Dave,

I didn't understand why the Mary argument is meant to involve negative possibility, while Zombies involve positive possibility. When I think about the Mary argument, (I try to) clearly and distinctly conceive of a scenario in which Mary knows all the physical facts, but can't work out the phenomenal facts. She knows all about red, she doesn't know what red is like. In those moments in which I find the argument compelling, I think that I have C and D conceived of just such a scenario. I don't see the asymmetry between the Zombie case and the Mary case. What am I missing here?

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Wed Mar 24 11:22:20 1999
Date: Wed, 24 Mar 1999 11:21:38 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Mary and Negative possibility
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Re Mary: It's true that one positively conceives of Mary, knowing all the physical facts but being unable to figure out the phenomenal facts. So Mary's situation is positively conceivable. But the question I was concerned with is whether "P and not-Q" (P = all the physical truths, Q = a phenomenal truth) is positively or negatively conceivable, as it's the conceivability or otherwise of this statement that hooks into the argument against materialism. Positively conceiving Mary isn't positively conceiving of a situation where P and not Q. Rather, the Mary scenario gives one good reason to believe that Q is not a priori deducible from P. That is, "P -> Q" is not a priori, i.e. "P and not-Q" is negatively conceivable.

I.e., what the Mary scenario suggests is that Mary can't a priori rule out the epistemic possibility that P is true and Q is false. That's not yet to say that such a scenario is positively conceivable. It may well be, and personally I think it is (e.g. via conceiving of zombies and inverted spectra), but that takes further argument, somewhat akin to the conceivability argument itself.

Re mathematics, Tim suggests interestingly that we say "3+3=6" is a priori precisely because we can't conceive otherwise. One subtlety here: presumably lots of complex mathematical statements M are such that we can't conceive otherwise, i.e. we can't positively conceive that not-M, but only because we can't positively conceive of either M or not-M. So perhaps the claim about "3+3=6" should rather be that in every situation of which we can positively conceive, "3+3=6" is verified. And the suggestion is that it's because "3+3=6" has this property that we say it's a priori.

Personally, I'm not certain of this. I agree that if " $3+3=6$ " was *not* verified in some positively conceivable scenario, that would be good reason to say it's not a priori. But I'm not sure that things work the other way. Maybe there can be support for apriority which is not grounded in trying to positively conceive otherwise. E.g., a mathematical proof seems to give support for a priority that doesn't come from trying to positively conceive otherwise. Of course one can still worry about the status of the axioms and inference rules here -- maybe those are so grounded? But this is far from obvious.

One worry is that if this were the only support for an a priority claim, such a claim would always be open to the objection that we haven't tried hard enough to conceive of relevant scenarios in which P is false, or that we're suffering from lack of imagination. Maybe that is a reasonable strategy for some apriority claims (e.g. Euclid's), but it's not clear that all apriority claims are vulnerable to this sort of objection. In particular, I think we have grounds for saying that " $3+3=6$ " is a priori without relying solely on our inability to conceive otherwise.

Maybe Tim is right that truth of P in all positively conceivable scenarios gives one reason to think P a priori. As such, it might only be defeasible reason, though. E.g. maybe Euclid did have good though defeasible reason to think Euclidean geometry a priori. Of course this eventually got defeated both by people positively conceiving of other geometries and by showing the formal consistency of other principles. It's not obvious to me that support for the a priority of all truths has this form, but it's an interesting claim.

Of course the fact that not-P is not *negatively* conceivable would be very good reason to say that P is a priori. But that's pretty trivial: to say that not-P is not negatively conceivable is to say that we can a priori rule out the epistemic possibility that not-P, which is more or less to say that we can a priori rule in P.

But it could be the case that P is true in all positively conceivable scenarios, while not-P is still negatively conceivable. E.g. because (1) we can't rule out the possibility of positively conceivable scenarios where P fails, or (2) maybe P is an odd sort of truth that is verified in *all* ideally positively conceivable scenarios, but still is not a priori. (My favorite example of the latter is "There are inconceivable features of the world.")

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Mar 24 12:56:08 1999

Date: Wed, 24 Mar 1999 13:54:16 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Josh Cowley <josh@MATH.ARIZONA.EDU>

Subject: Ideal conceivability

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

I'm still having problems with the notion of ideal conceivability (IC). A couple of the suggested definitions of IC are, "conceivable on ideal rational reflection," "conceivable after all possible reasoning is done," and perhaps "will remain conceivable given any more reasoning." Roughly my worry is that IC winds up being trivially identical with logical possibility. Here is how the argument goes. Being a non-ideal rational agent I don't know how to define ideal

rational reflection except as something like,

(A) "given some initial premises or concepts, reasoning which results in all and only logically possible conclusions."

I don't mean to say that we can't say anything about ideal rationality. We can give descriptions of what rationality for us is. And the more detailed this description gets the closer we will be to describing ideal rationality. We can even suggest the removal of some of our cognitive constraints such as the lack of time. But having limited rationality places limits on what we can ultimately say about ideal rationality. What we have to go to in the end is that ideal rationality is whatever achieves (A).

Is there anything wrong with this? It looks like it is. The current debate is whether conceivability implies or is a guide to logical possibility. I take it our argument is that ideal conceivability is a guide to logical possibility. Now we need only show that *secunda facie* conceivability is a good guide to ideal conceivability. But if ideal conceivability is not merely a guide, but is just defined by logical possibility, then all the arguments that the conceivable doesn't imply the possible can be used to say that the *secunda facie* conceivable doesn't imply the IC.

What is needed is some way of defining IC that doesn't ultimately rest on logical possibility.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Mar 15 23:35:08 1999

Date: Tue, 16 Mar 1999 00:33:42 -0700

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik A Herman <erikh@U.ARIZONA.EDU>

Subject: Re: Papers

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

I have a hypothetical scenario to propose regarding the bounds of ideal a prioricity (consider it actual if you have a difficult time believing it). (joke)

Consider an *a priori* genius, a true "head-case", that spends most of her time in deep meditation and spends little time gathering empirical information. My intuition is that she might eventually come to be able to deduce the physical structure of her brain. After all, all the fundamental physical stuff: space, time, and matter, is right there in her head-- so it might very well be retrievable introspectively (especially in her state of being relatively clear of empirical clutter).

In short, my point is that even in meditating with respect only to a *priori*/ a *priori* deducible concepts, we might have access to physical truths as we do mathematical ones; my worry is that this "objective viewing of the mind" looks a *posteriori* minus the senses.

My questions:

Does the head-case case seem like a logically possible case of ideal a prioricity?

If so, how would the *a priori* theorist deal with my concerns? For instance, what is fundamentally different from her internal monitoring of her brain, than if she stared at her hand? Are BOTH cases a *posteriori*?

Erik H.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Mar 25 10:40:06 1999
x-sender: agillies@pop.u.arizona.edu
Date: Thu, 25 Mar 1999 11:53:57 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: positive vs. negative conceivability
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

All,

I've been thinking about some ways of making the positive/negative conceivability (PC/NC) distinction clearer. Here are some thoughts on the matter.

First, NC is a bit trickier than we have been taking it to be. The official line is that something is NC iff there is no apparent contradiction in it. There aren't contradictions in *things*, so we have to be clearer about what we mean. One option is to say that a state of affairs S is NC iff the statement expressing S is apparently satisfiable. The trouble is that satisfiability requires us to picture a scenario (clearly and distinctly) which makes the statement expressing S come out true. But this is just what PC was supposed to be. So the satisfiability route won't work. The reason it won't work is instructive, though: any way of cashing out NC which is semantical will likely run into the same problem.

Maybe we can think of NC along *formal* lines. Say that a statement P is NC iff P is not apparently a contradiction, where "contradiction" is understood purely formally (as a matter of syntax, if you like). This won't quite do either: we need to add a patch to the effect that P is not apparently a contradiction, nor is it apparent that on the assumption of only P we can derive a contradiction. But the patch is in the same spirit: derivability is not a semantical consideration.

As for PC, we can treat that as a thorough-going semantical notion. A statement P is PC iff P is apparently satisfiable. Saying that P is (apparently) satisfiable requires us to picture a world in which P comes out as true. And this is what we want out of PC.

Drawing the lines in this way has some nice features. For one thing, it explains why NC seems easier to do. It's pretty hard to reason formally without relying on semantic considerations, and so it's easy to overlook formal contradictions (especially when they are derived somewhere down the line). In the Goldbach case, for instance, it is easier to pump intuitions to the effect that it is NC either way. But you can't pump any PC intuitions at all here: they aren't both satisfiable, not even apparently. Also, PC and NC can still be combined with the other dimensions of conceivability we talked about the other night.

Cheers,
Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Thu Mar 25 13:18:29 1999
Date: Thu, 25 Mar 1999 13:09:57 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: positive and negative conceivability, ideal conceivability

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Re Thony's comments on positive and negative conceivability. The first thing to note is that as defined, the various sorts of conceivability are all predicates of *statements* S. So S is negatively conceivable if there is no apparent contradiction in the hypothesis that S, or if one cannot rule out S a priori. Then one gets prima facie or ideal negative conceivability depending on whether one goes with surface reasoning or ideal reasoning. I think it's right to say that this is a semantic notion. Syntactic considerations re formal logic don't really play a part here. Even the talk of "contradiction" is not really essential; it's the inability to rule S out a priori that's central.

Re positive conceivability, "apparently satisfiable" is not a bad way to get at this, but of course that will come down to what "satisfiable" means here. If satisfiable means "possible", this comes down to something like van Cleve's definition of strong conceivability. One trouble there is that we want type-B materialists to come out saying zombies are conceivable but not possible; but on this definition, it's not clear that zombies will come out conceivable for a type-B materialist, as they will judge that zombies are impossible. Maybe there's a sense in which zombies are still "apparently possible" for the type-B materialist, but it's tricky to bring it out. That's why I go with clear and distinct conceptions instead (which is maybe more or less what Thony means by "picturing a world").

Re Josh's comment on ideal conceivability, I talked about this to some extent in our meeting yesterday. It's certainly true that defining "ideal reasoning" here as reasoning that tracks possibility will not be helpful, as it will render the conceivability/possibility thesis trivial. So we need some substantive notion of ideal reasoning.

Or at least, we need a substantive notion of better and worse reasoning. What's really needed is some kind of partial ordering on the space of reasoners, such that for a pair X and Y, it can be the case that X is "less limited" than Y, because X does not suffer from some cognitive limitation that Y suffers from. E.g., X has more memory or attention capacity than Y, or more concepts, or makes better inferences, and so on.

In addition to this, we also need the notion of "P is conceivable for X" (a sort of prima facie notion). Putting these two things together, we can say that P is ideally conceivable if there is some X such that P is conceivable for X, and for all Y that is less limited than X, P is conceivable for Y.

This does require the notion of a cognitive limitation and the corresponding partial ordering. I think we at least have a good intuitive grasp on this notion, though I don't claim it's 100% well defined. And I don't think it's defined in terms of logical possibility.

Still, maybe Josh's point that our own limitations limit what we can say about ideal rationality can be transferred here to the point that our limitations limit what we can say about cognitive limitations! (Maybe there are cognitive limitations we're unaware of, for example.) I'm not sure how much of a problem that is. As long as we have a good tacit grasp of the notion of a cognitive limitation, the notion of

ideal conceivability will be well-defined; what's required for the definition is that for X and Y, our statement "X is more limited than Y" can be true or false, irrespective of whether we can determine the truth or falsity. Of course this puts certain limits on our epistemic access to ideal conceivability, but we knew that already.

I suppose one might object that our notion of a cognitive limitation is so vague that for particularly advanced X and Y, our statement "X is more limited than Y" is neither true nor false, even though in some sense it "should" be true or false. I'm not sure exactly how coherent that is. At least, it seems that we can here defer to our smarter counterparts who can make the determination, so the truth of our statements of cognitive limitations is to some extent parasitic on their judgments. Or maybe one can just say that the very fact that the claim in question "should" be true or false suggests that our claim is true or false, it's just that we don't know it. (Like saying that certain mathematical statements are true/false even though we don't know it.) A tricky issue.

The strategy mentioned above also helps with the potential circularity problem. We want to connect possibility to ideal conceivability, but ideal conceiv*ability* is defined in terms of possibility (what various possible smarter beings can conceive). The conc-poss claim wasn't really intended as a reductive definition of possibility, so I'm not sure how much of a problem that is, but in any case, I'd like to think that any circle here is informative rather than vicious. Even if we defined ideal conceivability in terms of what certain conceivable beings could conceive, thus might be nontrivial, as long as we have facts about what we can conceive to get things off the ground. We can conceive of beings who lack some of our cognitive limitations, and they will be able to conceive of things that we can't conceive, among which things will be beings smarter than them, who will be able to conceive of things that they can't conceive, and so on. So maybe the circle here will be "expanding" and will eventually settle on a useful substantive notion of ideal conceivability.

At least, it would be nice if it did! But I can't claim that anything here is crystal clear. All thoughts are welcome.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Mar 26 13:59:11 1999
Date: Fri, 26 Mar 1999 14:58:00 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: weakly conceivable
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Van Cleve suggests that a proposition P is *weakly conceivable* for S iff S does not "see" that P is impossible. I realize that it is perhaps undesirable to include possibility in one's definition of conceivability. Nevertheless, I thought that weak conceivability sounds rather like negative conceivability. Van Cleve dismisses weak conceivability as a guide to possibility because of the Goldbach example: since we seem to be unable to see either that Goldbach's Conjecture is true or that it is false, we do not see either G or not-G as impossible. Thus, since we can weakly conceive of both and they are contradictory, weak conceivability cannot be a guide to possibility.

I do not really see how negative conceivability is different weak

conceivability. We could introduce the notion of ideal weak conceivability, which would presumably rule out the contradiction implied by the GC case-- presumably there is an ideal conceiver who can see that either G or not-G is impossible. Or, if by "seeing" we mean "clear and distinct perception", it does not seem clear that we can weakly conceive either G or not-G.

I think the point I am trying to make here is somewhat similar to Josh's-- it seems that, even though the definitions of the various types of conceivability do not explicitly refer to possibility, the notion of possibility is in some sense required to make sense of conceivability.

Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 28 13:10:00 1999
Date: Sun, 28 Mar 1999 13:09:48 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: weakly conceivable
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Anthony writes:

>Van Cleve suggests that a proposition P is *weakly conceivable* for S iff
>S does not "see" that P is impossible. I realize that it is perhaps
>undesirable to include possibility in one's definition of conceivability.
>Nevertheless, I thought that weak conceivability sounds rather like
>negative conceivability.

You're right that negative conceivability is a lot like van Cleve's weak conceivability (and positive conceivability is a lot like van Cleve's strong conceivability). I prefer to define it without making direct reference to possibility and impossibility: P is *prima facie* negatively conceivable if one cannot rule out P a priori, and P is ideally negatively conceivable if not-P is not a priori.

van Cleve's notion is obviously similar to this, at least if we restrict to the domain of primary possibility and conceivability. For P to be *prima facie* negatively conceivable (for S) is for S to rule P out a priori, which is for S to judge a priori that the actual world is a world where P is false, which is for S to judge a priori that all worlds considered as actual are worlds where P is false, which for S to judge a priori that P is 1-impossible, which is almost for P to be weakly conceivable for S. The main loophole is the chance that one might see that P is 1-impossible without judging this a priori; if so, P might be negatively but not weakly conceivable. That sort of loophole is common for 2-possibility, but rare for 1-possibility.

>Van Cleve dismisses weak conceivability as a
>guide to possibility because of the Goldbach example: since we seem to be
>unable to see either that Goldbach's Conjecture is true or that it is
>>false, we do not see either G or not-G as impossible. Thus, since we can
>weakly conceive of both and they are contradictory, weak conceivability
>cannot be a guide to possibility.
>

>I do not really see how negative conceivability is different weak
>conceivability. We could introduce the notion of ideal weak
>conceivability, which would presumably rule out the contradiction implied
>by the GC case-- presumably there is an ideal conceiver who can see that
>either G or not-G is impossible. Or, if by "seeing" we mean "clear and

>distinct perception", it does not seem clear that we can weakly conceive
>either G or not-G.

Right, weak conceivability isn't a good guide to possibility for the same reason that prima facie negative conceivability isn't (e.g. the cases you discuss above). Your two suggested corrections to the notion correspond to ideal negative conceivability and to prima facie positive conceivability respectively, both of which are much better guides to possibility.

>I think the point I am trying to make here is somewhat similar to Josh's--
>it seems that, even though the definitions of the various types of
>conceivability do not explicitly refer to possibility, the notion of
>possibility is in some sense required to make sense of conceivability.

Well, I'd like to think my way of defining things doesn't import the modal notion as directly as van Cleve's. As usual, though, modal notions may be playing a subtle role, especially in the idealized version (which is in effect about what **can** be known a priori by some possible being).

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 28 21:38:53 1999

Date: Sun, 28 Mar 1999 22:37:34 -0700

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Angela J Burnette <aburnett@U.ARIZONA.EDU>

Subject: Re: positive and negative conceivability, ideal conceivability

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Regarding Josh's comments about ideal conceivability and Dave's responses, there seems to be some slippage between thinking of the ideal conceiver as being a better reasoner in terms of further reasoning/cognitive powers and being a better reasoner in terms of different reasoning/ cognitive powers... Sometimes it has been asserted that the ideal conceiver is merely faster, older, and the like, but basically of the same kind, and at other times the ideal conceiver has had capabilities predicated of it that would seem to suggest that its powers of conceivability are of a different kind altogether, for example Brad's speculation that the ideal conceiver could possess the concepts of what it's like to be a bat (shouldn't such phenomenal concepts be intrinsic to the bat by definition?), or the idea that the ideal conceiver could be aware of its own cognitive limitations (it seems like it could be aware of ours, but not its own if it merely possesses further reasoning powers, beings that are like us should be cognitively closed to their own cog. limitations)... it seems to me that we can only legitimately abstract away from our own case if we are considering further reasoning which is basically of the same kind as our own... i.e. something like faster reasoning by a being with infinite time and memory... but if we are considering a conceiver who is better in virtue of being a different kind of reasoner, then I'm not sure we are warranted in our abstractions... an additional worry for me is that in order to get the move from conceivability to possibility off the ground we have to posit a conceiver who would not need to make that jump... i.e., it is unclear to me that something like what Josh has argued for isn't the case because positing the ideal conceiver is like collapsing the explanatory gap rather than bridging it... So, I would like some clarity on either why nothing said of the ideal conceiver makes him of a different kind than us, or, why this isn't a problem for abstracting away from our own case...

angela

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 28 22:48:20 1999
Date: Sun, 28 Mar 1999 23:45:35 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Re: ideal conception
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I have a concern similar to Angela's re the ideal conceiver. We need a notion of ideal rational reflection, but it seems that it must be of a kind that we can plausibly extend from our own cognitive powers. This would seem to preclude an ideal conceiver capable of understanding its own cognitive limitations or anything as miraculous as that.

We need a notion of "rational reflection" that is not different in kind than what we presently have (or any natural extension of that). Given this more modest extension of reasoning powers, however, there should be plausible counterexamples to the thesis from apriori reasoning to necessity. The Godel case, for one. This goes as follows. Given a minimum requirement for rationality--say, that any rational process could be described systematically or as a sequence of steps--we could not have an ideal conceiver solving any number-theoretic statements (Diophantine equations, say) without limit. This would place the notorious Godel formula outside the limits of ideal rational reflection. If, on the other hand, the ideal conceiver could have apriori access to the primary intensions of G statements, we would of course thwart these cases as real counterexamples, but at the expense of rendering the notion of "rational reflection" or more generally "thought" very vaguely defined, and of little help in the task of extending the scope of conceivability outward from our own case, for which, as Angela points out, we should have something of the same "kind".

So I think we've got, on the one hand, some potent counterexamples to theses like strong aprioricity and negpos, and on the other we've got a notion of "thought" or "rational reflection" that pulls too far away from our own case.

Erik

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 28 21:38:53 1999
Date: Sun, 28 Mar 1999 22:37:34 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Angela J Burnette <aburnett@U.ARIZONA.EDU>
Subject: Re: positive and negative conceivability, ideal conceivability
To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Regarding Josh's comments about ideal conceivability and Dave's responses, there seems to be some slippage between thinking of the ideal conceiver as being a better reasoner in terms of further reasoning/cognitive powers and being a better reasoner in terms of different reasoning/ cognitive powers... Sometimes it has been asserted that the ideal conceiver is merely faster, older, and the like, but basically of the same kind, and at other times the ideal conceiver has had capabilities predicated of it that would seem to suggest that its powers of conceivability are of a different kind altogether, for example Brad's speculation that the ideal conceiver could possess the concepts of what it's like to be a bat (shouldn't such phenomenal concepts be intrinsic to the bat by definition?), or the idea that the ideal conceiver could be aware of its own cognitive limitations (it seems like it could be aware of ours, but not its own if it merely possesses further reasoning powers, beings that are like us should be cognitively closed to their own cog. limitations)... it seems to me that we can only legitimately abstract away from our own case if we are considering further reasoning which is basically of the same kind as our own... i.e. something like faster reasoning by a being with infinite time and memory... but if we are considering a conceiver who is better in virtue of being a different kind of reasoner, then I'm not sure we are warranted in our abstractions... an additional worry for me is that in order to get the move from conceivability to possibility off the ground we have to posit a conceiver who would not need to make that jump... i.e., it is unclear to me that something like what Josh has argued for isn't the case because positing the ideal conceiver is like collapsing the explanatory gap rather than bridging it... So, I would like some clarity on either why nothing said of the ideal conceiver makes him of a different kind than us, or, why this isn't a problem for abstracting away from our own case...

angela

From owner-modality@LISTSERV.ARIZONA.EDU Sun Mar 28 22:48:20 1999

Date: Sun, 28 Mar 1999 23:45:35 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik J Larson <erikl@U.ARIZONA.EDU>

Subject: Re: ideal conception

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

I have a concern similar to Angela's re the ideal conceiver. We need a notion of ideal rational reflection, but it seems that it must be of a kind that we can plausibly extend from our own cognitive powers. This would seem to preclude an ideal conceiver capable of understanding its own cognitive limitations or anything as miraculous as that.

We need a notion of "rational reflection" that is not different in kind than what we presently have (or any natural extension of that). Given this more modest extension of reasoning powers, however, there should be plausible counterexamples to the thesis from apriori reasoning to necessity. The Godel case, for one. This goes as follows. Given a minimum requirement for rationality--say, that any rational process could be described systematically or as a sequence of steps--we could not have an ideal conceiver solving any number-theoretic statements (Diophantine equations, say) without limit. This would place the notorious Godel formula outside the limits of ideal rational reflection. If, on the other hand, the ideal conceiver could have apriori access to the primary

intensions of G statements, we would of course thwart these cases as real counterexamples, but at the expense of rendering the notion of "rational reflection" or more generally "thought" very vaguely defined, and of little help in the task of extending the scope of conceivability outward from our own case, for which, as Angela points out, we should have something of the same "kind".

So I think we've got, on the one hand, some potent counterexamples to theses like strong aprioricity and negpos, and on the other we've got a notion of "thought" or "rational reflection" that pulls too far away from our own case.

Erik

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Mon Mar 29 19:46:51 1999
Date: Mon, 29 Mar 1999 19:43:11 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: ideal conception
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Re Angela's and Erik's comments on ideal conceivability:

Angela writes:

>Regarding Josh's comments about ideal conceivability and Dave's responses,
>there seems to be some slippage between thinking of the ideal
>conceiver as being a better reasoner in terms of further
>reasoning/cognitive powers and being a better reasoner in terms of
>different reasoning/ cognitive powers...Sometimes it has been asserted
>that the ideal conceiver is merely faster, older, and the like, but
>basically of the same kind, and at other times the ideal conceiver has had
>capabilities predicated of it that would seem to suggest that its powers
>of conceivability are of a different kind altogether, for example Brad's
>speculation that the ideal conceiver could possess the concepts of what
>it's like to be a bat(shouldn't such phenomenal concepts be intrinsic to
>the bat by definition?), or the idea that the ideal conceiver could be
>aware of its own cognitive limitations (it seems like it could be aware
>of ours, but not its own if it merely possesses further reasoning
>powers, beings that are like us should be cognitively closed to their
>own cog. limitations)...

Well, it's true that we don't have an exhaustive list of the relevant cognitive limitations. It would be nice if we did, but at least so far, it isn't in sight. Still, what matters is that we have a determinate enough concept of limitation such that there's a fact of the matter about whether A is less limited than B (whether or not we now explicitly include that limitation in any list).

The case where B has concepts that A does not seems to be a straightforward case of that. So if B is just like us except that B has further concepts (further phenomenal concepts, say), then it seems reasonable to say that B is less limited than us in the relevant sense. So I'd like to think this isn't an unreasonable idealization. Indeed, in the standard definition of apriority, we say P is a priori if someone *who has the relevant concepts* can know P with experience-independent justification. On this definition it seems reasonable to say that there can be a priori relations among alien phenomenal concepts, even though we don't have those concepts.

You raise the interesting issue of whether any non-bat could have a concepts that can describe a bat's experience. This is tricky, but I'm not sure why it should be ruled out by definition. I can conceive of what it is like to be certain other humans (at least in certain respects, e.g. visually) even though I'm not them. And arguably I might conceive roughly of what it's like visually to be a dog, even though I'm not a dog. There's nothing contradictory about the idea that my phenomenal concepts might apply to another being's experience. And if I can do it with dogs, there's no obvious reason why some other being could do it with bats.

Problems would arise if it turns out that there are some phenomenal concepts that simply can't be had by other beings. Or perhaps, phenomenal concepts that simply can't be had by sophisticated beings. Just say experience E (perhaps the experience of stupidity?) could not be had by a sufficiently smart being, and further that such a being could not even form a direct concept of that sort of experience. If that were the case, then arguably the idealization here won't help in determining what's a priori or ideally conceivable re E (relatively ideal conceivers won't even have the concept!). That would be awfully tricky; but it's not obvious that such concepts exist.

If such concepts exist, they don't raise problems for the inference from ideal conceivability to possibility, but they might raise difficulties for the inference from possibility to ideal conceivability. These are problems in the same vicinity as the "open inconceivabilities" I mentioned last week. I think such inconceivabilities are probably the biggest obstacle to "pure modal rationalism"; but even the "strong modal rationalism" that survives without them is pretty strong. And of course, it is not obvious that such inconceivabilities really exist.

Re your second problem above: we are aware of some of our cognitive limitations, so why shouldn't a less limited being also be aware of theirs? And maybe they could even be aware of cognitive limitations they we haven't thought of yet? This is not to say that they can transcend those limitations, of course, any more than we can.

In both cases, I'd like to think that we're bringing in idealizations that are of a sort familiar from our own cases. We know some people have concepts that others lack, and we know that we're aware of some cognitive limitations. So I doesn't seem too great a stretch to shift to beings with concepts that we lack, and beings that are aware of cognitive limitations that we haven't thought about. Once we have the latter, we then have the "bootstrapping" possibility of removing those limitations in turn. So I'd like to think that this sort of idealizing can take us a fair way. Exactly where it takes us, of course, is not entirely obvious.

>it seems to me that we can only legitimately

>abstract away from
>our own case if we are considering further reasoning which is basically of
>the same kind as our own...i.e. something like faster reasoning by a being
>with infinite time and memory... but if
>we are considering a conceiver who is better in virtue of being a
>different kind of reasoner, then I'm not sure we are warranted in our
>abstractions...

The case of more concepts is like the cases of time and memory, I hope. Maybe the "second-order" idealizations (the ones we haven't thought of) go beyond this. And maybe that does raise a question about how well we can judge whether something is ideally conceivable or not. Of course there is a legitimate epistemic worry here: ideal conceivability is not epistemically transparent to us. But that's compatible with saying (i) there's a fact of the matter about whether P is ideally conceivable (whether or not we can know it), and (ii) in many cases, we can know whether or not P is ideally conceivable (e.g., we know it's ideally conceivable that something exists, and we know it's not ideally conceivable that $0+0=1$). So there's a fair amount we can know, even if things get hazy around the edges.

>an additional worry for me is that in order to get the move
>from conceivability to possibility off the ground we have to posit a
>conceiver who would not need to make that jump...i.e., it is unclear to me
>that something like what Josh has argued for isn't the case because
>positing the ideal conceiver is like collapsing the explanatory gap rather
>than bridging it...So, I would like some clarity on either why nothing
>said of the ideal conceiver makes him of a different kind than us, or, why
>this isn't a problem for abstracting away from our own case...

Hmm. I suppose the idea here is that (on the type-B position) a relatively ideal reasoner would just see that zombies are impossible, and would not be able to (positively or negatively) conceive of zombies? It's not obvious to me why this should be the case. Even the first half isn't obvious, but the second half is particularly unclear to me. What will rule out the clear and distinct conception of zombies for the ideal reasoner? Maybe something substantive in the a priori domain -- but then we just have type-A or type-C materialism, which is compatible with strong modal rationalism. If it's just that the reasoner knows that zombies are impossible, it seems to me that he/she would say, OK, zombies are ideally conceivable, they're just not possible. Presumably the ideal reasoner will be able to form just as clear and distinct a conception as we can, absent some substantive a priori obstacle.

So anyway, if this is right, it seems to me that we're not buying the conceivability/possibility link trivially. Someone like a type-B materialist who denies the link will still deny it on this way of putting things. Of course *if* there is a deep a priori contradiction in the notion of a zombie just that a relatively ideal reasoner will find it, then zombies will come out both conceivable and impossible, but that's as it should be.

(Maybe this does argue for not defining "positively conceivable" in terms of "seems possible", though. One could (could) argue that for the ideal reasoner, all and only the possible will seem possible, in which case the CP link will be trivial. This claim about the ideal reasoner is not obvious to me, but in any case, one can avoid the trivializing conclusion by going for a definition of positive conceivability that is not cast in terms of possibility. E.g., the clear and distinct conception idea.)

Let me know if I've missed some of the point here.

Erik writes:

>I have a concern similar to Angela's re the ideal conceiver. We need a
>notion of ideal rational reflection, but it seems that it must be of a
>kind that we can plausibly extend from our own cognitive powers. This
>would seem to preclude an ideal conceiver capable of understanding its own
>cognitive limitations or anything as miraculous as that.

Again, I'm not sure what's miraculous about understanding one's
cognitive limitations. Even we can do that!

>We need a notion of "rational reflection" that is not different in kind
>than what we presently have (or any natural extension of that). Given
>this more modest extension of reasoning powers, however, there should be
>plausible counterexamples to the thesis from apriori reasoning to
>necessity. The Godel case, for one. This goes as follows. Given a
>minimum requirement for rationality--say, that any rational process could
>be described systematically or as a sequence of steps--we could not have
>an ideal conceiver solving any number-theoretic statements (Diophantine
>equations, say) without limit. This would place the notorious Godel
>formula outside the limits of ideal rational reflection.

Well, even with the step-sequence restriction (a computability
restriction, let's say), any given statement of mathematics (our Godel
sentence G , say) could arguably still be knowable by *some* relatively
ideal reasoner. (This arguably follows from the considerations about
iterated Godelization that I previously attributed to Kleene, but which
Shaughan tells me should be attributed to Feferman.) Of course no
reasoner would know every truth, but every truth might be known by
some sufficiently sophisticated reasoner, which is all we need.

It's also not obvious why the computability restriction is essential.
I'd argue that we can form a pretty decent conception of e.g. a being
who can check out the truth of $P(n)$ for all n at once. The idea is
that when we can determine the truth of $P(n)$ for any given n , but
can't determine the truth of $\text{forall}(n) P(n)$, this is due to a
limitation on our part (e.g. seriality, or finite parallelness) that
can be idealized away from. Of course the idealized being here won't
be computable, but it's arguably still quite coherent. Though maybe
it will still induce queasiness.

>If, on the other
>hand, the ideal conceiver could have apriori access to the primary
>intensions of G statements, we would of course thwart these cases as real
>counterexamples, but at the expense of rendering the notion of "rational
>reflection" or more generally "thought" very vaguely defined, and of
>little help in the task of extending the scope of conceivability outward
>from our own case, for which, as Angela points out, we should have
>something of the same "kind".

Right. If we simply defined the ideal reasoner as being able to
determine the truth of any mathematical statement M , that would be by the
conclusion too cheaply. I'd like to think that neither of the methods
of idealization suggested above is quite as cheap as this, though. In
each case we're abstracting away from a clear cognitive limitation on
our part, such that we have at least a rough and ready conception of
what a being that does not suffer from that limitation might be like.

>So I think we've got, on the one hand, some potent counterexamples to
>theses like strong aprioricity and negpos, and on the other we've got a
>notion of "thought" or "rational reflection" that pulls too far away from
>our own case.

Right. I'd like to think that the idealizations suggested above give at least a potential "middle road". Of course the question is whether such a middle road will work across the board. If we characterize the reasoning powers too weakly, we get the counterexamples to scrutability, negpos, etc. If we characterize them too strongly, we buy the conclusion at cost of making it entirely trivial. I'd like to think the middle road above yields the desired conclusion while earning it nontrivially. The hope is that something like this will work across the board. But I don't have anything like a proof of that.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Mar 29 14:09:14 1999
Date: Mon, 29 Mar 1999 15:01:27 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Brad J Thompson <bradt@U.ARIZONA.EDU>
Subject: Re: Loar's mode of presentation argument
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

On Mon, 29 Mar 1999, Erik J Larson wrote:

> Maybe we should add some antecedent clause, to the effect that "If we have
> access to the correct primary intensions of phenomenal and physical
> states, and if we are able to completely specify (without fear of
> confusion or cognitive obscurities etc) these intensions...", then we get
> necessarily distinct primary intensions and a positively conceivable zombie
> scenario and a good foothold on the anti-materialist conclusion. But for
> myself, all I can see is the former argument-- we cannot see any necessary
> connection between physical and phenomenal states. And this is just a
> fact about what we can and can't see, and not much more. So I'm looking
> for someone to pull me out of the epistemic trap, so to speak (I'm
> guessing it will be Dave).

The first antecedent clause you mention seems to be required, but not the second one (the ability to completely specify the intensions). I assume that by "specify the intensions" you mean something like to be able to give a description. But notice that we haven't really done that for water, but instead use the phrase "watery stuff" as shorthand or to get a grip on the intension. Yet, we aren't troubled with the identification of water and H2O.

What is required is the assumption that our considerations of actual and counterfactual scenarios (conceiving of worlds as actual or counterfactual) engages our primary intensions and can be used as evidence for the nature of those intensions.

Your worry about getting out of an epistemic trap, though, is interesting. It seems to me that one can always raise worries about going from an epistemic conclusion to a metaphysical one--that follows straightforwardly from the fact that we can't rule out evil demons and the like. But if we *assume* that mad-dog skepticism is false, then I'm not sure we should be all that worried about the move from conceivability to possibility

(being careful in the ways we have discussed, that we are correctly describing the world that we are conceiving, etc.).

Brad J Thompson
bradt@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 30 01:30:36 1999
Date: Tue, 30 Mar 1999 00:23:21 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Loar's mode of presentation argument
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Re Loar, Angela writes:

>is he claiming that the conceivability of zombie worlds doesn't warrant
>the move from conceivability to possibility because we are fooled into
>thinking that they can actually come apart just because we can conceive of
>them coming apart (due to the fact that they are presented via different
>modes?) In other words, is he offering evidence to the effect
>that conc. doesn't nec. track possibility given the unique quality of
>phenomenal states? becaus it seems to me that the ideal conceiver would
>not be mistaken by different modes of presentation, but as this has not
>come up as a response to Loar (to my knowledge) I wonder if I am
>understanding the argument or not....

Right, I think that's what he's up to. He's trying to give reasons why even primary conceivability does not track primary possibility. He thinks that certain phenomenal and physical concepts can have different modes of presentation but still have the same primary intension across possible worlds, or as he puts it, "express the same property". If that were so -- i.e. if cognitively unconnected concepts could have the same primary intension -- he'd certainly have a way to block the epistemic/modal bridge. But I think his explanation ends up assuming what he needs to explain.

(N.B. This harks back to an earlier post of Tim's re Hill, which I haven't answered. Tim suggested that Hill didn't want to rely on a distinction between property and mode of presentation; so that on his view two cognitive distinct concepts might have the same mode of presentation. I'd be surprised if this were Hill's view, though. I think he thinks that the concepts in question have different modes of presentation; it's just that those modes express (not just refer to) the same property. It's almost definitional in much of philosophy that a mode of presentation is precisely what account for the cognitive distinctness of two concepts.

Tim's examples of tactile and visual experience of a property seem to involve a difference in both mode of presentation and in expressed properties. One concept fixes reference by "what causes this visual experience", one by "what causes this tactile experience"; different reference-fixing properties, even if the same property referred to. Of course if there are concepts like this which are a priori connected (a la Molyneux), then we'll have the same expressed property, but the thesis in question won't be endangered.)

Re ideal conceivers, I'm not sure why the two modes of presentation might not still be distinct even for them. Maybe the ideal reasoner

will know that the two concepts pick out the same property, just as Loar and others type-B materialists "know" this. But Loar and the others can still conceive of zombies, etc. So it seems to me that the ideal reasoner might equally still find zombies conceivable, even if he/she knows that they aren't metaphysically possible. To deny that would require an a priori connection between the concepts, and I don't think Loar thinks there is such a thing.

Maybe the ideal reasoner could close the gap *if* the epistemic distinctness of the concepts was due to a cognitive limitation on our part. But I don't think that's really what Loar thinks. He says explicitly somewhere that even ideal reasoning won't connect the concepts a priori. As before, I think that's the central distinction between the type-B and type-C materialists.

>The argument may be flawed in
>its specific guise. But I think there is a general confusion here, about
>exactly what the anti-materialist argument can sustain metaphysically
>given the epistemic starting point. I am tempted to say that the
>(anti-materialist) argument establishes this: Because we cannot see any
>necessary connection between physical and phenomenal states--the
>conditional $P \rightarrow Q$ is not apriori--we cannot positively establish the truth
>of materialism. I take it that the argument is stronger: Because
>physical and phenomenal states necessarily have different primary
>intensions, materialism is false (add in "there is a logically possible
>world where P holds but not Q and so on). The second formulation of the
>argument gets to the conclusion, but it seems that it rests on much
>shakier epistemic ground.

Well, the second is closer. I'd rather say something like "because physical and phenomenal concepts are not a priori connected, they have distinct primary intensions, so [via a chain of reasoning] materialism is false". Now, your objection is that the argument establishes only that we can't see the connection between the intensions. I can see two ways this might do. (1) We can't see an a priori connection, but a smarter being could see an a priori connection. (2) There's no a priori connection, but the two primary intensions are identical all the same. The first is the type-C materialist position; the second is the type-B position. The first preserves a deep epistemic/modal link, so perhaps you're pointing to the second.

What's wrong with the second? Well, my argument relies on one of the various epistemic/modal bridging principles (whether the strong a priority thesis, or the positive conceivability to possibility thesis, or ...), which seem to hold in all the usual cases. So denying them arguably would lead to a sui generis exception, which might be seen as ad hoc. And arguably this would make metaphysical modality brute and inexplicable. But the central reason against them I think is the argument against "modal dualism" that I outline toward the end of the PPR reply. We'll be talking about those considerations in more detail in a couple of weeks.

>Maybe we should add some antecedent clause, to the effect that "If we have
>access to the correct primary intensions of phenomenal and physical
>states, and if we are able to completely specify (without fear of
>confusion or cognitive obscurities etc) these intensions...", then we get
>necessarily distinct primary intensions and a positively conceivable zombie
>scenario and a good foothold on the anti-materialist conclusion. But for
>myself, all I can see is the former argument-- we cannot see any necessary
>connection between physical and phenomenal states. And this is just a
>fact about what we can and can't see, and not much more. So I'm looking

>for someone to pull me out of the epistemic trap, so to speak (I'm
>guessing it will be Dave).

Well, hopefully the above helps. Your point here is that if we don't have perfect a priori access to primary intensions, we can jump to false conclusions. And that seems right. Again, there seem to be two different sorts of gap. First, we haven't done enough (or good enough) a priori reasoning, so we think the two concepts have different PIs even though better reasoning would reveal they are the same. ("A=B" is a priori, we just haven't figured that out yet.) Second, the two concepts have the same PI even though that's not brought out by any amount of a priori reasoning ("A=B" is not a priori but is still 1-necessary). Type-C and type-B respectively, again. The former position keeps the epistemic/modal bridge but exploits our potential cognitive limitations. The second junks the bridge. I'm not sure which exactly you're pointing to here. On my view, the second is problematic because of modal dualism, and the first is problematic because it seems that the only remotely tenable way to support it is some sort of analytic functionalism, which seems to get the meaning of phenomenal concepts wrong.

The general question of a priori access to primary intensions is very interesting and important, though. It connects to the scrutability issue. Basically, some parts of the 2-D framework assume that for a concept C and a given world W, given a specification of W, we can a priori determine what the reference of C in W (considered as actual) is. Same for truth-value of statements S. That seems to plausibly fit many cases, the armchair methodology in the theory of reference, etc. But it does presuppose something like the generalized scrutability thesis: that given a complete qualitative description of the actual world (and of actual-world candidates), one can know the reference of our terms and the truth-value of our statements.

What is scrutability is false? E.g., just say the epistemic theory of vagueness is correct, so one can't know the truth-value of some "X is tall" given complete qualitative info. Or just say there are mathematical statements that are true and necessary but not a priori? Then in a certain sense, the primary intensions of the relevant terms and statements are not a priori accessible. E.g. the math statement might have its primary intension true in all worlds, but we couldn't know that. And something similar for the vague statement.

In these cases, I think one can say that the notion of a primary intension splits into two notions. There is the *fixing intension*, which goes with what the reference or truth-value in a given world would really be (if that world is actual). And there is the *epistemic intension*, which goes with the best a priori judgment about what the reference or truth-value in a given world will be.

In the vagueness case, the fixing intension of "X is tall" might be true or false in all worlds, while the epistemic intension is indeterminate in many of them. In the mathematical case, the fixing intension of P might be true in all worlds, while the epistemic intension is indeterminate. And so on.

I'd like to think that the scrutability thesis is true, so these cases can't arise. If so, the fixing intension and the epistemic intension are always the same. If such cases do arise, though, things get tricky.

What if an opponent suggests that the consciousness case is like this?

In fact, Joe Levine and Andrew Melnyk have both suggested that a loophole may arise from the assumption of accessibility of primary intensions (see Levine's review). I think that inaccessibility of PI may not be enough for the materialist here. In the cases above, this inaccessibility delivers inscrutable truths (gaps between negative conceivability and possibility). But zombies are positively conceivable, so the opponent needs strong necessities, not just inscrutable truths. And it's not clear that inaccessibility of PIs can deliver strong necessities. At the very least, it would need a much more radical form of inaccessibility (e.g. one that positively misleads us about a PI, rather than just not telling us everything).

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 30 09:55:14 1999

Date: Tue, 30 Mar 1999 10:53:22 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Rachael J Parkinson <rachaelp@U.ARIZONA.EDU>

Subject: Re: ideal conception

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

I want to echo some of Angela's concerns here, and particularly take issue with point ii) that Chalmers makes:

> The case of more concepts is like the cases of time and memory, I
> hope. Maybe the "second-order" idealizations (the ones we haven't
> thought of) go beyond this. And maybe that does raise a question
> about how well we can judge whether something is ideally conceivable
> or not. Of course there is a legitimate epistemic worry here: ideal
> conceivability is not epistemically transparent to us. But that's
> compatible with saying (i) there's a fact of the matter about whether
> P is ideally conceivable (whether or not we can know it), and (ii) in
> many cases, we can know whether or not P is ideally conceivable (e.g.,
> we know it's ideally conceivable that something exists, and we know
> it's not ideally conceivable that $0+0=1$). So there's a fair amount we
> can know, even if things get hazy around the edges.

Josh posed a concern about whether we have cognitive limitations that limit what we can say about our cognitive limitations. Dave responded by suggesting that we have a grasp of our own cognitive limitations which we can use as a springboard to get to an ideal conceiver. We can conceive of beings which lack our cognitive limitations, those beings can conceive of smarter beings which lack their cognitive limitations and so forth in an expanding circle. Hopefully, this will eventually give us a useful notion of an ideal conceiver.

But this seems problematic to me. Given that we may not be aware of our own cognitive limitations, it seems that we do not really have access to ideal conceivability. (point ii) How do we know that it is not conceivable that $0+0=1$ or that there are round squares? Perhaps we have a cognitive limitation that we do not come close to recognizing which prevents us from seeing that these things *are* ideally conceivable. If we do not have access to what an ideal conceiver can conceive of, this seems to seriously hamper our ability to do metaphysics, particularly when it comes to determining logically possible worlds.

-Rachael

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 30 14:31:45 1999

Date: Tue, 30 Mar 1999 14:24:44 -0800

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: ideal conception
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Rachael writes:

>But this seems problematic to me. Given that we may not be aware of our
>own cognitive limitations, it seems that we do not really have access to
>ideal conceivability. (point ii) How do we know that it is not conceivable
>that $0+0=1$ or that there are round squares? Perhaps we have a cognitive
>limitation that we do not come close to recognizing which prevents us
>from seeing that these things *are* ideally conceivable. If we do not have
>access to what an ideal conceiver can conceive of, this seems to
>seriously hamper our ability to do metaphysics, particularly when it comes
>to determining logically possible worlds.

It's true that our epistemic access to ideal conceivability is limited
by not knowing exactly what a relatively ideal conceiver could and
couldn't conceal. But I'd like to think it isn't this limited. I
think we *know* that $0+0=1$ and we know that this is a priori.
Similarly, I think we know that it is a priori that there are no round
squares. Given that we know that, we know that round squares are not
(positively or negatively) conceivable even for a relatively ideal
reasoner, and the same for $0+0=1$.

It seems to me that to be skeptical about these things is just to be
skeptical about whether we really know a priori that $0+0$ is not 1.
After all, if a smart enough being can conceive of a world such that
if that world is actual, $0+0=1$, then we can't know a priori that $0+0$
is not 1. Now maybe that's an interesting sort of skepticism, but I
don't think talk of ideal conceivability makes it more or less
plausible. I take it most of us think we have reasons to think
" $0+0=0$ " is true and a priori; so we have good reason to believe that
even an ideal conceiver couldn't conceive otherwise.

Similarly, I think we have reason to believe that "round squares do
not exist" is a priori, so we have good reason to believe that even
relatively ideal reasoners can't conceive of them. And we have good
reason to believe that "there is a phone on this desk" is a
posteriori, and that even a relatively ideal reasoner could conceive
otherwise. That's to say, we have good reason to believe that any
amount of further and better reasoning isn't going to change the
status of the knowledge here.

The moral, perhaps, is that even though we can't give an exhaustive
characterization of what makes for relatively ideal reasoning, that
doesn't mean we have no grip on what sort of things relatively ideal
reasoning can and can't tell us. To be skeptical about that, I think,
would be to be skeptical about our reasoning processes in general.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sat Apr 3 12:55:28 1999
Date: Sat, 3 Apr 1999 13:53:59 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: dogs
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

In Dave's response to Angela re the capabilities of the "ideal conceiver", he suggests that, "...Arguably, I can conceive roughly of what it's like visually to be a dog, even though I'm not a dog." This seems to be accurate. Whereas we have rod and cone structures in our eyes, dogs only have rod structures (I think..). Given that we have good reasons for thinking that cones are responsible for our experience of color, we think that dogs see in black and white. It seems quite straightforward to imagine the visual experiences of a dog-- they are probably somewhat like the experience of seeing a black and white movie.

It seems to me that it is somewhat more difficult to conceive of a dogs olfactory experiences. Avalanche rescue dogs are capable of quickly detecting the odor of people buried beneath considerable amounts of very hard snow (3 meters or more, sometimes). They are further able to distinguish the smell of the buried victim from the smell of the person handling them and other rescuers on the scene. I was once buried in a snowcave some six feet below the surface, and the rescue dog found me and had dug down to me in about 2 minutes.

I don't have the faintest idea what it would be like to be able to make the kinds of olfactory discriminations that dogs are capable of making. It does not seem to be me to be particularly likely that even an ideal conceiver would be able to do this. I seems fairly straightforward to imagine a what a dog's visula experiences are like because we assume that their visual experiences are somehow less rich than our own. But, given that dogs have olfactory capabilities far superior to ours, itjust doesn't seem plausible to suggest that any non-canine being, however idealized and with whatever instrumentation you please, could conceive of what a dogs olfactory experiences are like.

I'm not sure if all of this amounts to anything (except praise for the virtues of dogs).

Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Wed Apr 14 18:51:55 1999
Date: Wed, 14 Apr 1999 18:48:45 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: inscrutability, inconceivability, etc.
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Josh writes:

>Turning to yesterday's discussion, I'm unclear exactly what the
>difference between and inscrutable truth and an open inconceivability
>is. Both, I take it, are cases that are negatively conceivable but
>not positively conceivable. Both are possible. Perhaps open
>inconceivabilities are things that are to weird to have a concept of,
>but I don't think we want to use "weirdness" as a formal distinction
>(though I admit it is an informal distinction). So what did I miss?

Both open inconceivabilities and inscrutabilities fall into (or are closely related to) the "twilight zone" of statements that are negatively conceivable but not positively conceivable, but they are different. Informally, the difference is that open inconceivabilities (e.g., "there are inconceivable features of the world?") are verified as *false* in all complete positively conceivable scenarios, whereas inscrutabilities (e.g., some wondrous mathematical statement?) are *unsettled* in some complete positively conceivable scenarios.

I'll work through this and some other relevant things in more detail. To start, here are some of the definitions.

S is in the TWILIGHT ZONE if S is negatively conceivable but not positively conceivable. [N.B. "Ideal primary" is assumed throughout.]

NEGPOS PRINCIPLE: If S is negatively conceivable, S is positively conceivable. [I.e., there is nothing in the twilight zone.]

D is PC-COMPLETE if (a) D is positively conceivable, and (b) if D-and-F is positively conceivable, then D implies F.

Note: This is a slight variation on the definition of PC-completeness given before. It's straightforwardly equivalent to the previous definition, but is hopefully a little more intuitive. Also, recall that "A implies B" should always be read as "'A -> B' is a priori".

S is an INSCRUTABLE TRUTH if S is true and some PC-complete truth D does not imply S.

SCRUTABILITY OF TRUTH: If D is a PC-complete truth and S is true, then D implies S. [I.e., there are no inscrutable truths.]

Explanation: scrutability of truth says that all truths are implied by a complete qualitative description (i.e. a PC-complete description) of the actual world.

Next we generalize this to all possible worlds:

S is a GENERALIZED INSCRUTABLE if there exists some PC-complete D such that D-and-S is negatively conceivable, but D does not imply S.

D is NC-COMPLETE if (a) D is negatively conceivable, and (b) if D-and-F is negatively conceivable, then D implies F.

GENERALIZED SCRUTABILITY: If D is PC-complete, D is NC-complete.

Note: I didn't define generalized inscrutables, as opposed to generalized scrutability, in the lectures or notes, but the definition here is the natural one. Basically we take the definition of an inscrutable truth and relax the requirement that D hold of the actual world, and that S be true there.

Similarly, generalized inscrutability is the natural generalization of the inscrutability of truth. In effect, it says that for any possible world, a complete qualitative description of that world leaves nothing unsettled. It's easy to see that generalized scrutability as defined says precisely that there are no generalized inscrutabilities. (If this isn't obvious to you, working through it should be a worthwhile exercise.)

S is an OPEN INCONCEIVABILITY if S is negatively conceivable, but for all PC-complete D, D implies not-S.

NOINCONCEIVABILITY: No S is an open inconceivability.

It's straightforward to see that if S is an open inconceivability, it cannot be a generalized inscrutable, and vice versa. If S is an open inconceivability, it is verified as false in every PC-complete scenario, so it can't be left unsettled in any PC-complete scenario, which is what inscrutability requires. More formally: S is an open inconceivability $\leftrightarrow$ for all PC-complete D, D implies not-S $\leftrightarrow$ there is no PC-complete D such that D-and S is negatively conceivable $\rightarrow$ S is not a generalized inscrutable.

It's nevertheless true that both open inconceivabilities and generalized inscrutables, if they exist, would provide counterexamples to NEGPOS and inhabitants of the twilight zone. To see this:

(i) If S is a generalized inscrutable, then there is some PC-complete D such that D-and-S is negatively conceivable but D does not imply S. In this case D-and-S is negatively conceivable but not positively conceivable (if D-and-S were positively conceivable, D would not be PC-complete). So D-and-S is in the twilight zone.

(ii) If S is an open inconceivability, then S is negatively conceivable. But S is not positively conceivable: If S were positively conceivable, then there would be some PC-complete D such that D implies S.** So S is in the twilight zone.

[** This is a general principle about PC-completeness that is nontrivial but not hard to argue for. If S is positively conceivable, one can in effect construct D by conjoining as much as one can to S while still retaining positive conceivability. A maximal such conjunction will be PC-complete. The nontrivial part is making the case that there will always be a maximal conjunction here, as opposed to an ever-increasing series without a maximal element, but I think that case can plausibly be made. If the claim is doubted, we need to slightly modify the definition of open inconceivability.]

So, if NOINCONCEIVABILITY or GENERALIZED SCRUTABILITY is false, then NEGPOS is false.

The reverse also holds, as I said in class. Here's the proof I didn't have time to give. Just say NEGPOS is false, and there is some S that is negatively conceivable but not positively conceivable. Is S ruled out by all PC-complete D? If yes, then S is an open inconceivability, so NOINCONCEIVABILITY is false. If no, then there is some PC-complete D such that D does not imply not-S. But D cannot imply S (if it did, then S would be positively conceivable). So S is a generalized inscrutable, and GENERALIZED SCRUTABILITY is false.

All that is to say that NEGPOS is equivalent to NOINCONCEIVABILITY and GENERALIZED SCRUTABILITY. Or more informally, that the character of the twilight zone is exhausted by open inconceivabilities and generalized inscrutables.

All this provides a closer analysis of modal rationalism. Pure modal rationalism says that S is positively conceivable iff it is negatively conceivable iff it is possible. Weak modal rationalism says just that

if S is positively conceivable, it is possible (i.e. there are no strong necessities). Strong modal rationalism says that if S is negatively conceivable, it is possible (i.e. there is nothing in the twilight zone).

PURE MODAL RATIONALISM: pos con $\leftrightarrow$ neg con $\leftrightarrow$ pos
WEAK MODAL RATIONALISM: pos con $\rightarrow$ poss
STRONG MODAL RATIONALISM: neg con $\rightarrow$ poss
NEGPOS: neg con $\rightarrow$ pos con

It's not hard to see that pure modal rationalism is equivalent to the conjunction of weak modal rationalism and the negpos principle. Clearly, PMR entails WMR and NEGPOS. In the reverse direction, if WMR and NEGPOS hold, then neg con $\rightarrow$ pos con $\rightarrow$ poss. It's obvious that poss $\rightarrow$ neg con (if something is possible, it can't be ruled out a priori), so the circle is closed: neg con $\leftrightarrow$ pos con $\leftrightarrow$ poss.

So, pure modal rationalism is equivalent to the claim that there are no strong necessities and nothing in the twilight zone. This in turn is equivalent (via the previous discussion) that there are no strong necessities, no generalized inscrutables, and no open inconceivabilities. Hence these three problems exhaust the obstacles to pure modal rationalism.

Personally, I am fairly confident about ruling out strong necessities (and thus about weak modal rationalism), e.g. by the arguments we've discussed in class. I am less sure about ruling out generalized inscrutables, but I'd like to think there are none, and that there are at least some decent arguments for thinking so. And I'm less sure in turn about open inconceivabilities, which seem to be harder to get a grip on (as we found in class). If we were able to rule out all three, though, the resulting pure modal rationalism would certainly give a very clean and elegant picture of the epistemology of modality, and of the shape of the modal universe.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 15 10:34:12 1999
Date: Thu, 15 Apr 1999 10:14:42 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: Open inc's
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Dave specified that the only properties that we consider when talking about the possibility of open inconceivabilities are those that are non relational and nondispositional. I suppose that this rules out the case I'm about to make. At the moment, however, I'm not entirely clear how we can distinguish properties in this way.

Consider two universes, U1 and U2. Both contain only two particles, and in each the distance between the two particles is diminishing such that, at a certain point, the two particles will pass very close to each other and then continue off for a very long voyage into infinity, assuming no interaction between them. In U1, however, there is a force of gravity such that the two particles will actually start orbiting around each other. The two particles never get very close together, however, as, in each world, one of the particles spontaneously decays with an infinitesimal release of energy (and it's the same particle in each world that decays, if one can make sense of such a notion). U1 and U2 have identical histories and,

it seems, there is no possible way for any conceiver to tell the two worlds apart.

I'm assuming that not even an ideal conceiver will be able to discern differences in uninstantiated laws, as our ideal conceiver would then be, I presume, quite inconceivable to us. Now I assume my example is ruled out given that the only difference in U1 and U2 is in terms of a dispositional property. Presumably, a red thing that exists in this univers but which will not, at any time in its history, interact with any other matter (it spontaneously coalesced out of a little patch of energy, suppose) is still red.

It seems to me that in both cases, if it is meaningful to talk of properties at all, it is only insofar as a property has the potential for being discerned by some agent. But it is not clear to me why a property may then be ruled out simply because the event the event that would have instantiated that property never came to pass.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 15 18:30:05 1999

Date: Thu, 15 Apr 1999 18:25:29 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: Open inc's

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Anthony writes:

>Dave specified that the only properties that we consider when talking
>about the possibility of open inconceivabilities are those that are non
>relational and nondispositional. I suppose that this rules out the case
>I'm about to make. At the moment, however, I'm not entirely clear how we
>can distinguish properties in this way.

Actually, I didn't mean to rule out the possibility that dispositional or relational properties could be inconceivable. I was just pointing to intrinsic (nondispositional, nonrelational) properties as a particularly fruitful class where inconceivable properties might be found. But that's not to say that they might not be found elsewhere.

Anthony goes on to give the example of two universes with the same histories, but different uninstantiated laws, and suggests that the difference between these worlds might be inconceivable. The way I see things, both worlds are quite conceivable, and indeed Anthony did a nice job of describing them for us so we can conceive of each of them. What's true is that the difference between the worlds is not *perceivable* -- they will "look" the same to a being within those worlds, for example. But this just suggests that there is more to conceiving than perceiving. There are other examples of that: e.g. where one conceives of different theories compatible with the same empirical evidence (as e.g. in interpreting quantum mechanics), or where one conceives of unperceivable objects (e.g. a "peekaboo unicorn").

>It seems to me that in both cases, if it is meaningful to talk of
>properties at all, it is only insofar as a property has the potential for
>being discerned by some agent. But it is not clear to me why a property
>may then be ruled out simply because the event the event that would have
>instantiated that property never came to pass.

Hmm, I think many would question the first claim. One could argue

that in just this case, there's a real difference in properties of the worlds (a difference in laws) that would not be "discerned" by an agent in those worlds. I suppose one could argue that it is "discernible" in that there are some counterfactual states of affairs where the difference could show up. But in some cases (e.g. the unicorn or the theories) even this might not hold.

Nevertheless, in all these cases the properties can be *conceived* by some agent. I think that corresponds to the thesis some people were defending in class -- that it's not meaningful to talk about properties that can't be conceived by any agent. I'd be interested to see any arguments for that thesis.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Fri Apr 16 14:35:15 1999
Date: Fri, 16 Apr 1999 14:33:30 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: An inscrutable truth? Probably not.
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Here's an argument for an inscrutable truth. It's open to a number of objections - one of which I think is good.

(1) There are truths about bat experiences. There are true statements of the form "Being a bat is like. . .". Call these truths BE-truths.

(2) Bats have bat experiences, but they cannot believe any truths about bat experiences, because they lack the requisite concepts.

(3) You can only believe a truth about bat experiences if you yourself have battish (batty?) experiences.

(4) Nothing but a bat can have battish experiences.

Therefore,

(5) nothing can know BE truths.

In short: The only things that could be bats, but they can't, cos they don't have justified beliefs at all (or, at least, they don't have justified beliefs about their experiences.)

Some comments.

You might reject (1). I take it Dave (and Nagel) won't. They hold that there are intrinsic aspects of experience, that can only be known by having the experience.

You might reject (2). You might say that having an experience is not just necessary for having knowledge about that experience, but that it is also *sufficient*. We're in deep water here, but I think most would grant that you need more conceptual sophistication to know about your own experience than just the experience itself. You might reject (4). Why can't there be superbats? Superbats are just like bats - they have the same experiences as bats - but they have the conceptual sophistication to know about their own experiences. This is probably the weakest link in the argument. In fact, I guess I think that this objection is right.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Fri Apr 16 14:49:46 1999
Date: Fri, 16 Apr 1999 14:47:16 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: An inscrutable truth? Probably not.
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Tim's bat example is interesting. I should note that this isn't really being put forward as an instance of an inscrutable truth, but rather as an instance of an unknowable (perhaps even unbelievable) truth. It's important to note that these are not the same thing (though arguably inscrutable truths, if they exist, will be unknowable).

Of the categories we've been talking about, the truths in question are probably most relevant to open inconceivability than anything else. One might use Tim's argument to say: (i) to conceive of these truths, one needs to have battish experiences; (ii) only bats can have battish experiences; (iii) bats don't have the relevant concepts; so (iv) no possible being can conceive of these truths.

Tim's superbat objection would strike at premise (ii) here, and indeed that seems reasonable. One might try getting around this by talking about "dumb bat experiences", the sort that can only be had by an unintelligent bat. Superbats could never have these experiences, I presume! But maybe they could form the concepts some other way.

Another relevant route is to question (i) (which corresponds to Tim's (3)). Arguably one can conceive of some experiences without having had those exact experiences, but rather by having experiences that are relevantly related to them. Hume's "missing shade of blue" is an example. And maybe e.g. we could use our own experiences to form concepts of certain "proto-experiences" that compose them, and then recombine the protoexperiential concepts to form concepts of experiences that we've never had. Maybe that sort of thing could lead a being without bat experiences to form concepts of bat experiences (and even of "dumb bat experiences"). It's not obvious that this is possible, but it's not obvious that it's impossible, either.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sun Apr 18 22:59:45 1999
x-sender: agillies@pop.u.arizona.edu

Date: Sun, 18 Apr 1999 23:13:49 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: open inconceivability
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

All, some thoughts on open inconceivabilities--OI's--(and why I can't conceive of any).

The general strategy is to show that, on the assumption that S is negatively conceivable we can show that there is a PC-complete D such that $D \rightarrow S$. After doing that, we can look at problematic points of the argument.

So, let S be any (primarily ideally) negatively conceivable statement. Let R be an ideal rational agent. S is thus negatively conceivable by R in the limit. That is, there is some ordinal stage of R's reasoning σ , such that: R sees no contradiction in S at σ and R sees no contradiction in S at all ordinals $> \sigma$. Let D be the generalized conjunction of a maximal consistent subsets of the statements that R has marked as "conceivable" such that S is a conjunct of D. D is a complete description of a world, we just need to see that it is a *PC*-description. To see this, note that if R tags a statement as noncontradictory in the limit, then the statement is noncontradictory. That is, it has a satisfying structure. So D has a satisfying structure. But surely D must then be PC-complete. $D \rightarrow S$, and so we're done.

Two things worth pointing out:

(1) Most of the work above is being done by tacit appeal to the equivalence I see between NI-in-the-limit to PC: if we are serious about full rationality in the limit (to the point of fixed points of ordinal updates), then this seems to me to be plausible. What else could we mean by NC in this case other than something like satisfiable?

(2) We need to explain away claim to OI status that statements like "There are inconceivable features of the world" might seem to have. I think that's pretty easy. First, they equivocate on "inconceivable"--sure there might be features which are inconceivable to us, but that's no surprise. And that doesn't establish that there are OI's. What's relevant is whether there are statements which are in the limit NC but not PC (and we're nowhere near the limit). But, if a statement is in-the-limit NC iff it is PC, then this is not possible.

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Mon Apr 19 00:11:21 1999
Date: Sun, 18 Apr 1999 23:59:53 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: open inconceivability
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Thony puts forward a general argument against the existence of open inconceivabilities. In fact he gives much more: an argument that

anything that is ideally negatively conceivable is also positively conceivable. If successful, such an argument would establish the negpos principle, and rule out both open inconceivabilities and inscrutable truths.

>So, let S be any (primarily ideally) negatively conceivable statement.
>Let R be an ideal rational agent. S is thus negatively conceivable by R
>in the limit. That is, there is some ordinal stage of R's reasoning
> σ , such that: R sees no contradiction in S at σ and R sees no
>contradiction in S at all ordinals $> \sigma$. Let D be the generalized
>conjunction of a maximal consistent subsets of the statements that R has
>marked as "conceivable" such that S is a conjunct of D. D is a complete
>description of a world, we just need to see that it is a
>*PC*-description. To see this, note that if R tags a statement as
>noncontradictory in the limit, then the statement is noncontradictory.
>That is, it has a satisfying structure. So D has a satisfying structure.
>But surely D must then be PC-complete. D $\rightarrow$ S, and so we're done.

Hmm. I'm not certain what you mean by "consistent" above. I think you mean something like "not contradictory". If so, then there will plausibly be such a D, and the D in question will be NC-complete. The crucial question, as you say, is whether D is also PC-complete. If D is PC, it will be PC-complete. But what reason is there to believe that D is PC?

Your argument turns on a "satisfying structure". I need to know just what a "satisfying structure" is. I see about three possibilities:

- (1) If a satisfying structure is merely a model that gives an interpretation for the terms such that D comes out true in that model, then there will certainly be a satisfying structure, but it's not at all obvious why the existence of such a structure implies positive conceivability, given that e.g. such a model may assign an interpretation of the terms in D quite different from our ordinary interpretation.
- (2) If a satisfying structure is a model such that the interpretation of D in the model is determined by the ordinary semantics of the terms in D, then it's less obvious that such a model will exist. If one is loose enough about what counts as preserving the ordinary interpretation of the terms, perhaps there will be a model, but in this case the inference from the existence of a model to the positive conceivability of D is again far from clear.
- (3) If a satisfying structure is just a possible world in which D is true, or even a positively conceivably situation in which D is true, then it's not obvious why such a satisfying structure should exist. Or at least, that's just what's at issue.

To clarify, you might show in detail how the argument would run for (a) D = the conjunction of all the physical truths about the world with some purported inscrutable truth (e.g. a statement about tallness, on the epistemic theory; or a mathematical statement that is true but not a priori), and (b) a purported open inconceivability such as "there are inconceivable features of the world". As it stands your argument seems to rule these things out perhaps just a little too easily, but maybe the details will help.

>Two things worth pointing out:
>(1) Most of the work above is being done by tacit appeal to the
>equivalence I see between NI-in-the-limit to PC: if we are serious about
>full rationality in the limit (to the point of fixed points of ordinal

>updates), then this seems to me to be plausible. What else could we mean
>by NC in this case other than something like satisfiable?

Well, NC in the limit means absence of any contradiction detectable through reasoning. That definition doesn't make any obvious claims about satisfiability. Certain weak forms of satisfiability may follow, but these don't obviously guarantee positive conceivability. And for strong forms of satisfiability that guarantee positive conceivability, it's not obvious why satisfiability in these senses follows from NC-completeness. At least, such a link is just what's at issue. But any argument for the link would be gratefully received!

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Apr 19 02:38:34 1999

X-Priority: 3

X-MSMail-Priority: Normal

X-MimeOLE: Produced By Microsoft MimeOLE V4.72.3110.3

Date: Mon, 19 Apr 1999 02:36:15 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Brad Thompson <bradt@U.ARIZONA.EDU>

Subject: A problem for the notion of a single ideal conceiver

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: R

The case Tim mentioned about "dumb bat experiences" makes me think of a related possible area of open inconceivabilities. Actually, I want to present a case in which there are truths which can be conceived of by some being but not by an ideal conceiver. I'm not sure whether such truths should be called open inconceivabilities. Such a case might require us to either 1) revise our notion of conceivability such that it doesn't depend on a single ideal conceiver but rather on being conceivable by *some* being, or 2) reject modal rationalism. I actually prefer 1. (I can't recall whether the notion of a single ideal conceiver was crucial in the first place. We seem to have presupposed it in discussion.)

OK, so the case goes like this. It seems that there *could* be a being whose conscious experience is holistically tied to its concepts. This is like an extreme case of Tim's bat case. The idea is that when one of these beings acquires a new concept, all of its conscious experiences are modified (slightly or radically, however you like). Assuming that having a concept of this being's phenomenal states requires having that state, then *any* being which has different concepts (and certainly an ideal conceiver which has all possible concepts) cannot have a concept of these experiences. Actually, as the preceding sentence shows, there is no being which can have all possible concepts.

Here is a similar case. Imagine a being whose conscious experiences are holistically tied to its *beliefs* rather than its concepts. Any change in belief changes all the qualitative feels of its experiences. So such a being who believes that P will have perhaps radically different types of phenomenal experiences from an identical being that believes that not-P or believes neither P nor not-P. Now, surely our ideal conceiver believes P, not-P, or has neither belief. If it believes P, then it can't have the experience of a being (in the hypothetical species) that believes not-P or neither, etc....

Now surely the beings that I've imagined are possible. I think that even human beings have some of these properties. In the conceptual case, imagine the difference between seeing a chessboard and lacking the concept of chess and seeing a chessboard *as* a chessboard. In the belief case, imagine the

difference between eating cowbrain and not knowing what you are eating and eating the brain and knowing it (yuk!).

I don't think we want to call the cases I've discussed open inconceivabilities. After all, they are conceivable by *some* being, just not by the ideal conceiver we've hypothesized. Do we need a single ideal conceiver? It seems like we don't, though I haven't thought this out. Things seem to get pretty messy without the notion of a single ideal conceiver. But I suppose we already had a gap between human conceivability and ideal conceivability. Now we have a gap between *any* being and "total conceivability"--that is, between what any particular being can conceive and the totality of things conceivable by at least one being.

Brad

-----Original Message-----

From: David Chalmers <chalmers@LING.UCSC.EDU>

To: MODALITY@LISTSERV.ARIZONA.EDU <MODALITY@LISTSERV.ARIZONA.EDU>

Date: Friday, April 16, 1999 2:46 PM

Subject: Re: An inscrutable truth? Probably not.

>Tim's bat example is interesting. I should note that this isn't
>really being put forward as an instance of an inscrutable truth, but
>rather as an instance of an unknowable (perhaps even unbelievable)
>truth. It's important to note that these are not the same thing
>(though arguably inscrutable truths, if they exist, will be
>unknowable).

>

>Of the categories we've been talking about, the truths in question are
>probably most relevant to open inconceivability than anything else.
>One might use Tim's argument to say: (i) to conceive of these truths,
>one needs to have battish experiences; (ii) only bats can have battish
>experiences; (iii) bats don't have the relevant concepts; so (iv) no
>possible being can conceive of these truths.

>

>Tim's superbat objection would strike at premise (ii) here, and indeed
>that seems reasonable. One might try getting around this by talking
>about "dumb bat experiences", the sort that can only be had by an
>unintelligent bat. Superbats could never have these experiences, I
>presume! But maybe they could form the concepts some other way.

>

>Another relevant route is to question (i) (which corresponds to Tim's
>(3)). Arguably one can conceive of some experiences without having
>had those exact experiences, but rather by having experiences that are
>relevantly related to them. Hume's "missing shade of blue" is an
>example. And maybe e.g. we could use our own experiences to form
>concepts of certain "proto-experiences" that compose them, and then
>recombine the protoexperiential concepts to form concepts of
>experiences that we've never had. Maybe that sort of thing could lead
>a being without bat experiences to form concepts of bat experiences
>(and even of "dumb bat experiences"). It's not obvious that this is
>possible, but it's not obvious that it's impossible, either.

>

>--Dave.

>

From owner-modality@LISTSERV.ARIZONA.EDU Mon Apr 19 21:14:06 1999

Date: Mon, 19 Apr 1999 21:13:48 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: A problem for the notion of a single ideal conceiver
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Brad discusses some problems for the notion of an ideal conceiver. I note that we decided that this notion might be problematic and was unnecessary: instead of defining ideal conceivability in terms of what an ideal conceiver can conceive, we can define it in terms of conceivability "in the limit" among non-ideal conceivers who get smarter and smarter (e.g., P is ideally conceivable if there is some X such that P is conceivable for X and P is conceivable for all beings less limited than X). That being said, the problem still raises issues even for the modified view of ideal conceivability.

>OK, so the case goes like this. It seems that there **could** be a being
>whose conscious experience is holistically tied to its concepts. This is
>like an extreme case of Tim's bat case. The idea is that when one of these
>beings acquires a new concept, all of its conscious experiences are modified
>(slightly or radically, however you like). Assuming that having a concept
>of this being's phenomenal states requires having that state, then **any**
>being which has different concepts (and certainly an ideal conceiver which
>has all possible concepts) cannot have a concept of these experiences.
>Actually, as the preceding sentence shows, there is no being which can have
>all possible concepts.

>
>Here is a similar case. Imagine a being whose conscious experiences are
>holistically tied to its **beliefs** rather than its concepts. Any change in
>belief changes all the qualitative feels of its experiences. So such a
>being who believes that P will have perhaps radically different types of
>phenomenal experiences from an identical being that believes that not-P or
>believes neither P nor not-P. Now, surely our ideal conceiver believes P,
>not-P, or has neither belief. If it believes P, then it can't have the
>experience of a being (in the hypothetical species) that believes not-P or
>neither, etc....

Interesting cases. These seem to be somewhat similar to the "dumb bat experiences", in that they are experiences that can only be had by beings below a certain level of sophistication. As such, then none of our "sufficiently smart" beings will be able to have those experiences. Does it then follow that these experiences are not ideally conceivable?

That conclusion seems to follow, **unless** there is a way the beings in question can conceive of the experiences without actually having them. We've seen that this can happen at least in some cases, e.g. Hume's "missing shade of blue", and novel experiences that involve recombination of elements of familiar experiences. Perhaps something like that could happen here. E.g., perhaps a smart being with a rich conceptual system could still conceive of the phenomenology of a being with a limited conceptual system, perhaps by "scaling down" their own phenomenology in their imagination, or by some sort of analogical conception. Similarly, a being that believes P can arguably conceive what it's like to believe that not-P, perhaps by "changing the polarity" of their current belief in the imagination. E.g., even though I believe I was born in Australia, I think I can conceive of the phenomenology of believing I wasn't born in Australia, even though I don't have that phenomenology myself.

Of course it isn't obvious that these methods will work for every possible case. But it's not obvious that they won't. If they do, things are OK. If they don't, it would cause trouble for the concept

of ideal conceivability, and certainly for the link between ideal conceivability and possibility. The possibilities in question would come out to be open inconceivabilities by the official definition, which might seem odd.

An alternative would be to create a new class for statements that are conceivable by some beings but not by smarter beings; but then the question is how to distinguish between the "coherent" cases like this (e.g. the ones just discussed) from the "incoherent" ones that are implicitly contradictory. That's particularly problematic if there are cases (a) involving concepts that can only be had by limited beings and (b) that involve subtle incoherencies on the part of those beings. We want to class these with the "incoherent" cases, but clearly the criterion of being undermined by rational reflection won't do, as beings that could perform the rational reflection wouldn't have the concepts! Maybe there is an alternative way of running the counterfactuals -- e.g. they're incoherent because *if* a being could have the concepts and be smarter, they would uncover a contradiction. Such a counterfactual would have an impossible antecedent and might thus be considered problematic, but arguably counterfactuals with impossible antecedents can sometimes have truth values.

In any case, this would clearly be a bit of a mess. So let's hope that such cases can't really arise! I.e., let's hope that the experiences in question can at least be conceived of by smarter beings, even if they can't be had by smarter beings.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 20 11:16:33 1999
Date: Tue, 20 Apr 1999 11:14:32 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik A Herman <erikh@U.ARIZONA.EDU>
Subject: ideal concievers
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

I think Brad's example of "a being whose conscious experiences are holistically tied to it's beliefs" is exactly the case with humans in the actual world. A change in belief DOES change the qualitative feel of things--and in a new way. As far as the ideal conceiver goes though, being ideal and all, can't it conceive of what it's like to beleive not-p while believing p? The ideal conceiver has considered both of them already and decided that p is the case, but probably "remembers" what it was like to belive not-p. So, using Brad's example, the ideal conceiver would be able to conceive of how you feel when you're eating cow brain in the case where you know what it is AND the case where you don't. Of course, there is going to be a different qualitative feel to the ideal conceiver because it can't remove itself wholly from being an ideal conceiver as opposed to the person actually EATING cow brain, but we would call what it is doing, "remembering", to be a form of "concieving of": [what it is like to eat cow brain without knowing that it is cow brain] wouldn't we?

erik h.

PS (I hope it is OK to refer to the ideal conceiver as "it")

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 20 12:33:45 1999
Date: Tue, 20 Apr 1999 12:12:12 -0700
Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: open conceivabilities or the lack there of
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Here is an argument that I think shows there are no open conceivabilities. First some of Dave's definitions:

- > S is an INSCRUTABLE TRUTH if S is true and some PC-complete truth D does not imply S.
- > S is a GENERALIZED INSCRUTABLE if there exists some PC-complete D such that D-and-S is negatively conceivable, but D does not imply S.
- > S is an OPEN INCONCEIVABILITY if S is negatively conceivable, but for all PC-complete D, D implies not-S.

Because of the truth constraint in the def of INSCRUTABLE TRUTH I'm going to focus on a comparison of generalized inscrutable (GI) and open inconceivability (OI).

The primary difference between GI's and OI's is that for GI's no D implies S while for OI's all Ds imply ~S. However, S is negatively conceivable for both GI and OI. What I'm not getting is how something can be negatively conceivable and yet be false in every conceivable world. Let's say that I know S is false in some PC-complete D. This is the same as saying that if I add S to D then I get a contradiction. Now if S implies a contradiction in every PC-complete world I can imagine then it seems to me that S leads to an obvious contradiction. But that is to deny that S is negatively conceivable.

Now if S isn't even negatively conceivable then I see no reason for thinking S is possible. On the other hand if S doesn't imply a contradiction in every world then there is some D such that S is either true or undecided. In the latter case we just have an inscrutable truth.

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 20 15:15:20 1999
Date: Tue, 20 Apr 1999 15:11:14 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: open conceivabilities or the lack there of
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Reply Josh writes:

- >Here is an argument that I think shows there are no open conceivabilities. First some of Dave's definitions:
 - >
 - >> S is an INSCRUTABLE TRUTH if S is true and some PC-complete truth D does not imply S.
 - >
 - >> S is a GENERALIZED INSCRUTABLE if there exists some PC-complete D such that D-and-S is negatively conceivable, but D does not imply S.
 - >
 - >> S is an OPEN INCONCEIVABILITY if S is negatively conceivable, but for all PC-complete D, D implies not-S.

>
>Because of the truth constraint in the def of INSCRUTABLE TRUTH I'm
>going to focus on a comparison of generalized inscrutable (GI) and open
>inconceivability (OI).
>
>The primary difference between GI's and OI's is that for GI's no D
>implies S while for OI's all Ds imply ~S. However, S is negatively
>conceivable for both GI and OI. What I'm not getting is how something
>can be negatively conceivable and yet be false in every conceivable
>world. Let's say that I know S is false in some PC-complete D. This
>is the same as saying that if I add S to D then I get a
>contradiction. Now if S implies a contradiction in every PC-complete
>world I can imagine then it seems to me that S leads to an obvious
>contradiction. But that is to deny that S is negatively conceivable.

Hmm. Let's take a statement such as "there are inconceivable features of the world". It's true that this is false in any PC-complete D, and combined with any such D leads to a contradiction. But that's not to say that S alone leads to a contradiction, or that S is negatively conceivable. That would require that not-S is a priori. Now *if* one could know a priori that there is some PC-complete description of the actual world, then it would more or less follow from the above that one could know a priori that not-S. But it's not obvious that we can know the antecedent a priori, and indeed that's what's at issue. So, it seems that for this argument to rule out open inconceivabilities, we'll first need a way to rule out open inconceivabilities!

>Now if S isn't even negatively conceivable then I see no reason for
>thinking S is possible. On the other hand if S doesn't imply a
>contradiction in every world then there is some D such that S is
>either true or undecided. In the latter case we just have an
>inscrutable truth.

Well, for an open inconceivability S, there is the chance that (i) S may imply a contradiction in all positively conceivable worlds, but (ii) S may not imply a contradiction in all possible worlds. If there are open inconceivabilities, there may be more possible worlds than positively conceivable worlds.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 20 14:55:55 1999
Date: Tue, 20 Apr 1999 14:53:12 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Rachael J Parkinson <rachaelp@U.ARIZONA.EDU>
Subject: Re: A problem for the notion of a single ideal conceiver
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Dave responded to Tim and Brad's cases by suggesting that beings can conceive of experiences without really having them. Hume's "missing shade of blue" may be such a case, and yet it seems to be much more simple than the "dumb bat" cases we've been discussing. I am not sure that it is possible for an ideal conceiver to conceive of what it's like to be a bat simply by scaling down his own phenomenological experiences. This is because the difference in phenomenological feel between an ideal conceiver and a bat seems to be a difference of kind and not of degree. I can not conceive of what it's like to be an ant by simply imagining myself, only much stupider.

Also, even if an ideal conceiver were to conceive of what it's like to be a bat by simply scaling down his own phenomenology, it seems that he would still have a lot of extra baggage in virtue of being an ideal conceiver that would taint his conception. I can remember what it's like to be five, but I don't know that I really have a pure concept of the feel of what it's like to be five because I have so much 'experience and wisdom' that lurks in the background when I try to conceive of it. (I can never really conceive of the wonder I felt toward the world because now I can better explain the world... Of course I could just take the little wonder that I feel now and multiply it, but I'm not sure that that really does the trick.)

I guess a lot rides on what is required to have a positive conception of another's experience- my intuitions seem to be that you do need to have the experience in order to conceive of it. (Of course, a person who has never had a headache can presumably conceive of what it's like to have a headache, but that brings up an epistemological problem- unless he's had the headache- how does he know that he's getting it right? Likewise for the ideal conceiver and the dumb bat.)

-Rachael

> That conclusion seems to follow, *unless* there is a way the beings in
> question can conceive of the experiences without actually having them.
> We've seen that this can happen at least in some cases, e.g. Hume's
> "missing shade of blue", and novel experiences that involve
> recombination of elements of familiar experiences. Perhaps something
> like that could happen here. E.g., perhaps a smart being with a rich
> conceptual system could still conceive of the phenomenology of a being
> with a limited conceptual system, perhaps by "scaling down" their own
> phenomenology in their imagination, or by some sort of analogical
> conception. Similarly, a being that believes P can arguably conceive
> what it's like to believe that not-P, perhaps by "changing the
> polarity" of their current belief in the imagination. E.g., even
> though I believe I was born in Australia, I think I can conceive of
> the phenomenology of believing I wasn't born in Australia, even though
> I don't have that phenomenology myself.
>

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 30 14:54:23 1999
Date: Tue, 30 Mar 1999 14:57:45 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Horgan on properties
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

A short note on Horgan and properties. . .

Horgan asks what information the Demon needs in order to complete the task of cosmic hermeneutics. The demon knows:

- (1) All the a priori accessible truths;
- (2) All the microphysical facts.

What else does it need? Horgan suggests that identity statements might do the trick. If all the macrophysical objects (states of affairs, etc.) are (ontologically, rather than conceptually) identical to various concatenations of microphysical objects, then macrophysical states of affairs necessarily supervene on microphysical states of affairs, *without the introduction of brute metaphysical relations,* which Horgan finds problematic.

But, says Horgan, there's a problem. "It is most implausible, on the face of it, to suppose that every sortal and non-sortal *property* expressible in our higher-level vocabulary is also finitely expressible in microphysical vocabulary. Surely nobody seriously believes that there is a single finitely-long microphysical predicate which expresses the property of being a chair, or being a coffee pot, or thinking of Vienna" (23).

I think that the above is true, but I (very tentatively) suggest that there may be around this particular problem. My suggestion is that the materialist go nominalist on properties, (as many of them are wont to do anyway). There ain't no such thing as the property of being a chair, or being a coffee pot, and so on. There are, of course, chairs, but 'there are chairs' is true not because of a property of chairhood (conceived of as some universal over and above the microphysical structure of the world), but because certain concrete particulars (tropes) stand in certain relations to each other. Now surely (some) relations are included in the basic microphysical facts that the demon knows - at least facts about the relations that microphysical particles bear to one another. The nominalist about properties would suggest that those property facts such as there are logically fall out of (are entailed by) the microphysical facts (where these include microphysical relations). My materialistic nominalist might agree that she can't define <chair> in reductive terms, but she can reduce the (finite) chair tropes, such as there are, to (1) and (2). Insofar as there are properties, they are fully ontologically reducible. Now, it may be that there are independent problems with this suggestion - it may not get you anywhere near cosmic hermeneutics - but that's another story.

PS. I don't find nominalism about properties attractive, but many do.

Well, it's a thought.

t.

Timothy J. Bayne
RM. 213 Social Science

Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Tue Mar 30 23:22:20 1999
Date: Tue, 30 Mar 1999 23:18:27 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Horgan on properties
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Tim writes:

>Horgan asks what information the Demon needs in order to complete the task
>of cosmic hermeneutics. The demon knows:
>(1) All the a priori accessible truths;
>(2) All the microphysical facts.
>
>What else does it need?

Pardon me for using this to get into a digression from Tim's point, to
address some general points re Horgan's framework and some things that
came up in class today.

Of course on my view, nothing else is needed. Or at least, (1) and
(2) are enough if we add (a) an indexical truth (saying where the
demon is now), (b) phenomenal truths, and (c) a that's-all truth. But
leaving those "complications" aside, it's not clear why we need to add
more.

In the dialectic as Horgan sets it up, he first grants the demon
knowledge of (2), then asks what else is needed to get to macro
knowledge. Clearly, *some* sort of bridge is required to get to macro
concepts. He canvases laws (no good), identities (no good), before
settling on "meaning constraints". This is a little different from
(1) above: as Horgan understands meaning constraints, they include a
posteriori constraints such as the constraint that "water" picks out
H₂O. One might argue that this supplementation of the demon's
physical knowledge base with ineliminably a posteriori knowledge
weakens its claim to be engaging in "cosmic hermeneutics". But in any
case I think that the demon doesn't need a posteriori meaning
constraints. A priori meaning constraints (the sort that correspond
to primary intensions) are enough for the demon to figure out the
macro truths in these cases. I think Horgan now accepts this and
regrets giving in to a posteriori meaning constraints here.

Further, it's arguably best not to put things in terms of giving the
demon "meaning constraints" at all (even the a priori sort), if these
are understood a further class of truths. E.g. it's not clear what
set of explicit a priori truths about "knowledge" will enable the
demon to give the right verdict on all the Gettier-et-al cases.
Rather, I think it's best just to give the demon the concept of
knowledge, and argue that armed with microphysical facts, that
concept, and sufficient a priori reasoning, the demon can figure out
the knowledge facts. More generally, I think that one just needs to
stipulate that the demon has the relevant concepts. Explicit "meaning
constraints" aren't necessary here and are a bit of a distraction. Of

course it may be that in giving the demon the concept we're giving it implicit meaning constraints, in some sense, but we needn't build that into the picture directly, and we certainly needn't give the demon any special truths directly in its inference base (as (1) above might suggest).

Horgan's different way of doing things relates to his "naturalized semantics" approach that someone (Brad?) mentioned in class. In effect, he argues that meaning constraints aren't really problematic, as they're just another class of high-level fact. Eg. the demon can figure out English truths by looking at English speakers, figuring out what their terms mean, and seeing whether or not their statements are true. And can presumably figure out meaning constraints the same way, including explicit truths about what 'water' in English refers to, etc..

I find this picture a little hard to follow, though. First, if we are really getting the meaning constraints "for free" via naturalized semantics, why bother building them in in the first place? The answer, presumably, is that even in the naturalized semantic case, we need *some* bridge to get us from microphysics to what is "true in English" and so on -- after all, we'll be stating such claims using such non-microphysical concepts as "true". So presumably, we'll at least need some bridge re the concepts of "truth", "reference", etc, otherwise we'll never be able even to get to claims about what is true in English, etc, even on the naturalized semantics approach. And presumably the bridge that is needed will be precisely something like meaning constraints for the concepts of "truth", "reference", etc. So it seems to me that Horgan can't really get the meaning constraints as cheaply as he'd like: he has to *start* with at least some of them.

I don't think starting with these meaning constraints is problematic myself (one just has to suppose that the demon has the concepts of truth, reference, etc), but it does suggest that the naturalized semantics approach can't do all Horgan's work for him. In essence, we have to suppose that the demon *has* and *uses* some of the relevant concepts, and doesn't just *mention* them as part of a third-person interpretation. At least, it has to have the concepts of truth, reference, etc. And once we've done that, we might as well give the demon arbitrary concepts. That way, to see if any truth T is derivable by cosmic hermeneutics, we just give the demon the concepts in T, physical knowledge, and let things rip. No need to do everything resolutely in the third-person, restricting the demon to "X's utterance of T is true" and such. The demon can just tell us, T!

The two approaches might even come apart in some cases. E.g. just say nonmental truths are a priori entailed by physical truths, but mental truths aren't. And now consider "water is wet". On the "concept use" approach to cosmic hermeneutics, where we give the demon (say, an idealized version of me) the concepts, then it will be able to figure out from microphysics etc that water is wet. On the "naturalized semantics" approach, though, the demon may not be able to figure out what "water" in English refers to, as that will probably depend on various mental facts about the intentions and beliefs, etc, of English speakers, which aren't accessible by hypothesis. If so, it will come out that cosmic hermeneutics doesn't even work for "water is wet", despite a priori derivability. I think that's the wrong result, and we want an understanding of CH so that CH is possible for "water is wet" in this case. If so, I think we need the concept-use approach.

>What else does it need? Horgan suggests that identity statements might do
>the trick. If all the macrophysical objects (states of affairs, etc.) are
>(ontologically, rather than conceptually) identical to various
>concatinations of microphysical objects, then macrophysical states of
>affairs necessarily supervene on microphysical states of affairs, *without
>the introduction of brute metaphysical relations,* which Horgan finds
>problematic.

Well, I'd argue that bringing in identities as basic here may well
bring in something "brute" in some sense. Something like the type-B
materialist bringing in a brute fact that pain = C-fibre firing, or
some such. If there is no epistemic link from microphysics to that,
then we have at least an epistemic bruteness, if not a metaphysical
bruteness. And it would seem that as with a posteriori meaning
constraints, we'd have then given up on pure cosmic hermeneutics.
But set that aside.

>But, says Horgan, there's a problem. "It is most implausible, on the face
>of it, to suppose that every sortal and non-sortal *property* expressible
>in our higher-level vocabulary is also finitely expressible in
>microphysical vocabulary. Surely nobody seriously believes that that there
>is a single finitely-long microphysical predicate which expresses the
>property of being a chair, or being a coffee pot, or thinking of Vienna"
>(23).

>
>I think that the above is true, but I (very tentatively) suggest that
>there may be around this particular problem. My suggestion is that the
>materialist go nominalist on properties, (as many of them are wont to do
>anyway). There ain't no such thing as the property of being a chair, or
>being a coffee pot, and so on. There are, of course, chairs, but 'there
>are chairs' is true not because of a property of chairhood (conceived of
>as some universal over and above the microphysical structure of the
>world), but because certain concrete particulars (tropes) stand in certain
>relations to each other. Now surely (some) relations are included in the
>basic microphysical facts that the demon knows - at least facts about the
>relations that microphysical particles bear to one another. The nominalist
>about properties would suggest that those property facts such as there
>are logically fall out of (are entailed by) the microphysical facts (where
>these include microphysical relations). My materialistic nominalist
>might agree that she can't define <chair> in reductive terms, but she can
>reduce the (finite) chair tropes, such as there are, to (1) and (2). Insofar
>as there are properties, they are fully ontologically reducible. Now, it
>may be that there are independent problems with this suggestion - it may
>not get you anywhere near cosmic hermeneutics - but that's another story.

Hmm, interesting. I'm not sure I completely follow. What are the
"bridges" that we are adding to get us from microphysics to macro
truths? Horgan canvased the possibility of adding identities such as
"water = H₂O", but the trouble was that analogous identities don't
seem available in many other cases (chairs, coffee pots, thinking).
You're suggesting maybe adding token reductions? I presume that would
be something like "this P₁ is a chair", "this P₂ is a chair", and so
on for every chair in the world? I guess the trouble, as you suggest,
is that this doesn't look much like cosmic hermeneutics. Rather, it's
more like taking microphysics and tagging every chair explicitly as a
chair, every coffee pot explicitly as a coffee pot, and so on. Not
much surprise that we can get to the chair facts from there!

I think your suggestion here might end up being relevant to a
metaphysical reduction, as opposed to the sort of epistemic reduction

Horgan is after. It might be that a token-token but not type-type identity is enough to save physicalism as a metaphysical doctrine (though there are many tricky issues here); and a tropewise reduction of properties might do much the same thing. But in building in so many references to specific instances, it seems to build in too much to qualify as cosmic hermeneutics. Rather, this might lead to a way in which physicalism could be true *without* cosmic hermeneutics being possible! That would be interesting, though of course at the end of the day I think such a combination isn't tenable. If there are these tropewise identifications, there ought to be some way of figuring out the identifications themselves a priori from the physical facts.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Apr 5 10:49:02 1999

x-sender: agillies@pop.u.arizona.edu

Date: Mon, 5 Apr 1999 11:01:48 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Anthony S Gillies <agillies@U.ARIZONA.EDU>

Subject: cosmic hermeneutics

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

All,

Some thoughts on cosmic hermeneutics (CH) and reductive explanation (RE).

There seems to be a bit of carelessness in talk of the possibility of CH with respect to what informaton we are to give a Laplacian-type knower: is it all the physical facts, or is it all he facts about physics? Byrne, e.g., seems to run the two together. But surely these are are conceptually distinct sets of facts, and their difference hinges on where we put natural laws. I think there is a dilemma lurking nearby, so I want to try to draw it out. (I'm not sure what force the dilemma might have for particular theories, over and above being a useful guide to taxonomy.)

Let F be the set of facts we give to an ideal knower, and let P be the conjunction of all the microphysical facts, and let L be the conjunction of all the (true) physical laws. By hypothesis P is in F. The question is whether L is also in F. If L is *not* in F, then I think we are getting into Byrne's worry about conceivability. It is hard to see how one can get from low-level sentences describing mass, charge, spin and so on to high-level descriptions about juiciness or whatever without using th sort of bridge laws that are conjuncts of L. And if that's right, then it isn't very clear in what sense it is a priori that F implies a given macro-event M. (And, if I read Josh's worry re Aristotle correctly, this is similar in spirit to his point.)

Now suppose that L *is* in F. So we give the ideal knower both all the micro-facts and all the causal laws. But then CH is possible iff the deductive-nomological theory of explanation is right. Right to Left: If the D-N model is right, then since explanation is just a deduction from conditions and laws to empirical facts, clearly CH is possible---CH amounts to "large scale" D-N prediction/explanation. Left to Right: If CH is possible, there is a deduction from F to M. But $F = \{P, L\}$, where P is the description of all the physical conditions (antecedent circumstances) and L is the statement of all the general laws. M is just a macro empirical fact. But these are the sufficient ingredients for the D-N model of explanation.

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 6 11:45:14 1999
x-sender: agillies@pop.u.arizona.edu
Date: Tue, 6 Apr 1999 11:29:38 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: Re: cosmic hermeneutics
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Well, I guess I mean *I* don't want to draw any strong conclusions, since I'm not sure that I'm convinced that D-N explanation is so bad (Salmon and his ilk notwithstanding). I might be convinced, and just don't know it, though.

Thony

>Thony, Just a point of clarification. I take it that at least part of
>your point about the D-N model is that, if we have committed ourselves to
>the D-N model to hold that CH is possible, this is potentially problematic
>(disastrous?) to the CH thesis, since there are well-known problems with
>D-N as an adequate model of scientific explanation (e.g., Salmon's point
>that D-N is neither necessary nor sufficient for explanation). (I realize
>that this is an extrapolation, since you say that there *aren't* any
>particularly strong conclusions here*, but I think it's clear that a
>commitment to a D-N model of explanation would be bad, so I suppose I
>would want to add that, if you are correct, there IS a strong conclusion
>here; namely, there is something amiss with the CH thesis on this score.)
>
>Erik
>(p.s. there is no way out for you, if you do not answer me I will hunt you
>down for further clarification)
>
>

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 6 10:53:23 1999
Date: Tue, 6 Apr 1999 10:49:24 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: ON what Cosmic Hermeneutics involves
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Clarificatory point. I thought that the CH claim is that a being with (1) all the physical facts; and (2) all a priori knowledge; can (3) come to know all the facts, where (3) should be read as, 'can at one time know everything there is to know'. But in fact Byrne nor, as far as I can tell, Horgan, puts the CH conclusion in this way. Rather, they say that for *any* truth, one can come to know it, not that the Laplacean demon can

come to know all the truths at once (although Horgan seems to gesture in this direction in at least one place). Call the former claim the existential claim, and the latter the universal claim.

The problem is this: as far as I can see, the Demon might know everything, but can it know that it knows everything? We, as theorists, can stipulate the 'that's all' fact for any world, but how does one establish the that's all truth from *within* a world? One might believe that one knows everything, and it might even be true that one knows everything, but how could the belief that one knows everything be justified? (Why hasn't this objection been levelled against theistic omnipotence? Because God is said to *create* everything, and thus (it seems) God knows when God knows everything. I think that's the reason.)

As I said, I think that Byrne (and perhaps) Horgan reject the universal claim in favor of the existential claim, and the existential claim isn't open to this problem. Or is it? Consider the claim: 'The demon knows everything'. The demon will (I guess) believe this, but how can it know it? But is this a truth? No. Is it a truth that the demon has true beliefs about everything? That may be true. Can it know that it has true beliefs about everything (i.e. can it know that it believes every truth). ? Again, I don't think so: how could you be justified that your set of beliefs is exhaustive? So, it seems that there is a particular truth that the demon cannot *know*.

Here's an objection I tried to develop, but the wheels fell off. If someone can put them on again, go ahead! Suppose a weird world, in which p is true iff the demon doesn't believe it. The fact of believing p causes p to be true no longer, and the fact of not-believing p causes p to be true. Now, if you put the demon outside of the world (as Horgan does) you might be able to say that, necessarily, there are no such facts p. But - as far as I can see - putting the demon inside the world seems to open up the possibility of such facts.

Problem is, fully developing this objection involves saying something about the relation of p and the physical facts. It needs to be the case that the demon can know the physical facts on which p supervenes, otherwise this isn't a counter-example to CH. And if looks like if the demon does know the physical facts on which p supervenes, and it knows all the a priori facts, then it will be able to work out whether or not p is true. In other words, the defender of CH seems to be within her rights in claiming that the notion of a p truth is incoherent.

Tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 6 14:46:54 1999
Date: Tue, 6 Apr 1999 14:42:27 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: cosmic hermeneutics
To: MODALITY@LISTSERV.ARIZONA.EDU

Status: R0

Re cosmic hermeneutics and D-N explanation, Thony writes:

>There seems to be a bit of carelessness in talk of the possibility of CH
>with respect to what informaton we are to give a Laplacian-type knower:
>is it all the physical facts, or is it all he facts about physics?
>Byrne, e.g., seems to run the two together. But surely these are are
>conceptually distinct sets of facts, and their difference hinges on where
>we put natural laws. I think there is a dilemma lurking nearby, so I
>want to try to draw it out. (I'm not sure what force the dilemma might
>have for particular theories, over and above being a useful guide to
>taxonomy.)

I think we are supposed to give the demon (a) all the microphysical facts and (b) all the microphysical laws. Just (a) will be enough if one is a Humean about laws and causation (in that case (b) will follow from (a) a priori). On a non-Humean views, one will need both, to bring in facts about laws, causation, dispositions, counterfactuals, etc.

>
>Let F be the set of facts we give to an ideal knower, and let P be the
>conjunction of all the microphysical facts, and let L be the conjunction
>of all the (true) physical laws. By hypothesis P is in F. The question
>is whether L is also in F. If L is *not* in F, then I think we are
>getting into Byrne's worry about conceivability. It is hard to see how
>one can get from low-level sentences describing mass, charge, spin and so
>on to high-level descriptions about juiciness or whatever without using
>th sort of bridge laws that are conjuncts of L. And if that's right,
>then it isn't very clear in what sense it is a priori that F implies a
>given macro-event M.

It's not clear to me whether you intend L to be the collection of microphysical laws, or of physical laws more generally (including macrophysical laws). In the first case, L will be in F, as suggested above. In the second case, L will not be in F, but the advocate of cosmic hermeneutics will argue that L follows from F a priori.

How do we get from microphysical facts and laws to facts about juiciness, etc? Presumably, by virtue of possessing the concept of juiciness. The CH claim is that armed with this concept, one grasps how this situation applies to various epistemic possibilities, and in particular, that one can know how it applies to a possibility microphysically described. So e.g. one will be able to get from a structural/functional characterization of an apple to knowledge of whether it is juicy. E.g., look to see how much water it contains! Where we've antecedently figured out what is water in the world, by the "watery stuff" method. And the CH advocate will argue that it is not a priori conceivable that something satisfies this physical characterization without being juicy. Etc.

Of course whether this is really possible for concepts such a juiciness is just what the debate over CH is about. I am personally agnostic about whether it is possible, though I'm sympathetic with the idea. What I do think is that the juiciness facts follow a priori from the physical plus phenomenal facts. Once one allows the demon to have phenomenal facts too, e.g. about what the apple looks like, then its task is much easier.

>Now suppose that L *is* in F. So we give the ideal knower both all the
>micro-facts and all the causal laws. But then CH is possible iff the

>deductive-nomological theory of explanation is right. Right to Left: If
>the D-N model is right, then since explanation is just a deduction from
>conditions and laws to empirical facts, clearly CH is possible---CH
>amounts to "large scale" D-N prediction/explanation. Left to Right: If
>CH is possible, there is a deduction from F to M. But F= {P, L}, where P
>is the description of all the physical conditions (antecedent
>circumstances) and L is the statement of all the general laws. M is just
>a macro empirical fact. But these are the sufficient ingredients for the
>D-N model of explanation.

Well, the laws most relevant to micro-macro D-N explanation are
micro-macro bridge laws. And we certainly won't be giving the demon
knowledge of micro-macro bridge laws. So L as construed here won't be
in F. With knowledge of bridge laws, the demon's task would be much
easier, but (as Horgan notes at one point) it trivializes things. It
will turn out that CH is true even on a property dualist view of
qualia, for example. (Even on my view of qualia, there are bridge
laws!) So CH would certainly not imply reductive explanation or any
such thing. For CH to be interesting, the demon needs to be able to
figure things out from microphysical facts and laws alone, in
combination with its possession of the high-level concept.

As Tim notes, there are various problems with the D-N model of
explanation. But I think most of those can be set aside here. The
most central point is that it is widely agreed now that D-N makes a
terrible model of *reductive* explanation (although Ernest Nagel
originally intended it this way). In helping itself to bridge laws,
D-N makes the reductive project relatively trivial, not obviously
reductive, and consistent with the falsity of materialism, as the
qualia example suggests. So most people now think reductive
explanation requires a stronger constraint.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 6 14:56:47 1999
Date: Tue, 6 Apr 1999 10:12:15 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Re: cosmic hermeneutics
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Thony, Just a point of clarification. I take it that at least part of
your point about the D-N model is that, if we have committed ourselves to
the D-N model to hold that CH is possible, this is potentially problematic
(disastrous?) to the CH thesis, since there are well-known problems with
D-N as an adequate model of scientific explanation (e.g., Salmon's point
that D-N is neither necessary or sufficient for explanation). (I realize
that this is an extrapolation, since you say that there *aren't* any
particularly strong conclusion here*, but I think it's clear that a
commitment to a D-N model of explanation would be bad, so I suppose I
would want to add that, if you are correct, there IS a strong conclusion
here; namely, there is something amiss with the CH thesis on this score.)

Erik
(p.s. there is no way out for you, if you do not answer me I will hunt you
down for further clarification)

On Mon, 5 Apr 1999, Anthony S Gillies wrote:

<http://www.u.arizona.edu/~chalmers/class/596b/week10.txt> (11 of 27) [4/7/2002 1:49:06 PM]

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 6 15:25:27 1999
Date: Tue, 6 Apr 1999 15:23:54 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: ON what Cosmic Hermeneutics involves
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Tim writes:

>Clarificatory point. I thought that the CH claim is that a being with (1)
>all the physical facts; and (2) all a priori knowledge; can (3) come to
>know all the facts, where (3) should be read as, 'can at one time know
>everything there is to know'. But in fact Byrne nor, as far as I can tell,
>Horgan, puts the CH conclusion in this way. Rather, they say that for
>*any* truth, one can come to know it, not that the Laplacean demon can
>come to know all the truths at once (although Horgan seems to gesture in
>this direction in at least one place). Call the former claim the
>existential claim, and the latter the universal claim.

I think it's probably better to interpret CH as making at most the
existential claim. The universal claim may have problems. E.g. there
is a "paradox of the knower" due to Montague and Kaplan that raises
Godel-style problems with the idea of a being who knows all truths.

One might argue that the existential claim entails some version of the
universal claim, since the conjunction of all truths is itself a
truth, so the existential claim applies to it. In response, a CH
advocate might suggest (1) that they are only dealing with finite
claims (it's not clear whether or not this would be ad hoc), or at
least (2) that the notion of the conjunction of all truths is not well
defined. Alternatively, they could put forward a slightly different
version of CH (see below).

>The problem is this: as far as I can see, the Demon might know everything,
>but can it know that it knows everything? We, as theorists, can stipulate
>the 'that's all' fact for any world, but how does one establish the that's
>all truth from *within* a world? One might believe that one knows
>everything, and it might even be true that one knows everything, but how
>could the belief that one knows everything be justified? (Why hasn't this
>objection been levelled against theistic omnipotence? Because God is said
>to *create* everything, and thus (it seems) God knows when God knows
>everything. I think that's the reason.)

Well, I take it that the "that's all" claim, like microphysics, is
simply stipulated to be in the demon's knowledge base. One could
argue about just how a demon could really know that, just as one could
argue about just how a demon could really know all the microphysical
truths, but that's not really the central issue here.

It's probably better to put the CH claim in terms of knowledge of
conditionals. For any truth Q, "P+T+I -> Q" is knowable a priori,
where P is the conjunction of the microphysical truths, T is the
that's-all truth, and I is the indexical truth.

>As I said, I think that Byrne (and perhaps) Horgan reject the universal
>claim in favor of the existential claim, and the existential claim isn't

>open to this problem. Or is it? Consider the claim: 'The demon knows
>everything'. The demon will (I guess) believe this, but how can it know
>it? But is this a truth? No. Is it a truth that the demon has true beliefs
>about everything? That may be true. Can it know that it has true beliefs
>about everything (i.e. can it know that it believes every truth). ? Again,
>I don't think so: how could you be justified that your set of beliefs is
>exhaustive? So, it seems that there is a particular truth that the demon
>cannot *know*.

Well, *if* the demon has the that's-all fact in the antecedent part of
its knowledge base, this obstacle will be removed. Again, you can
raise questions about whether a demon could really have that
antecedent knowledge, but we are just stipulating it here. I also
note that "the demon knows everything" will likely actually be
false, at least for the existential version of CH, as here the demon
knows truths one at a time, not all at once. Maybe the demon will
believe "everything is knowable", although there are tricky issues
even there.

>Here's an objection I tried to develop, but the wheels fell off. If
>someone can put them on again, go ahead! Suppose a weird world, in which
>p is true iff the demon doesn't believe it. The fact of believing p
>causes p to be true no longer, and the fact of not-believing p causes p to
>be true. Now, if you put the demon outside of the world (as Horgan does)
>you might be able to say that, necessarily, there are no such facts p. But
>- as far as I can see - putting the demon inside the world seems to open
>up the possibility of such facts.

Nice objection. In fact there is a perfectly straightforward such P.
Let's assume that the demon doesn't in fact know all truths (we're
going with the existential rather than the universal version, after
all). So there is some truth Q that the demon doesn't know. So let R
= "Q and the demon does not know that Q". R is clearly true. But if
the demon were to come to know that R, he would thereby come to know
that Q, so P would be false. Hence the demon cannot know that R!

This is precisely the "paradox of knowability" that we discussed in
class. There is a bit of literature on it. It clearly poses a
problem for some formulations of CH. It can't be the case that all
truths are knowable for the demon, and it can't even be the case that
given full physical knowledge, all truths are knowable.

The best response, I think, is to go back to the conditional
formulation, which says that the *conditional* from microphysics to
any given truth (i.e. "P+T+I -> Q" for any Q) is knowable a priori.
In the case where Q is the paradoxical R above, the demon can't know
that R, but it can still know a priori that *if* P+T+I, then R. Of
course, on the hypothesis of P+T+I, the demon will be able to figure
out both that Q above is true and that he doesn't know it, so he'll be
able to figure out R. Because the antecedent here is merely
hypothetical, this doesn't imply that he actually knows R, but he can
still figure out the conditional a priori.

>Problem is, fully developing this objection involves saying something
>about the relation of p and the physical facts. It needs to be the case
>that the demon can know the physical facts on which p supervenes,
>otherwise this isn't a counter-example to CH. And it looks like if the
>demon does know the physical facts on which p supervenes, and it knows all
>the a prior facts, then it will be able to work out whether or not p is
>true. In other words, the defender of CH seems to be within her rights in
>claiming that the notion of a p truth is incoherent.

I think your P is actually coherent, as above. In the case where the demon *knows* all the microphysical facts, it will be able to figure out Q, and "I didn't know that Q", but it coming to know these things it will render "Q and I don't know that Q" false, so the original truth was still arguably unknowable. But if we put things in the conditional format, there doesn't seem to be a problem with the demon knowing a priori that *if* microphysics is like so, then P is true, even though P itself is unknowable.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sat Apr 3 17:11:19 1999
Date: Sat, 3 Apr 1999 18:10:14 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: reductiv concepts
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Cosmic Hermenudics.

I want to push the Aristotle problem I was suggesting in class. The basic idea is that we have a person who has all the concepts of Aristotle (ie. no modern science) and is an ideal reasoner. We then give him all the micro-physical truths and the "that's all" fact. Now consider Aristotle's favorite wooden chair. The question I have is whether Aristotle can get from the micro-physical truths to "There are macro objects in the world." Without any further concepts. In particular whether he needs reductive concepts linking macro objects to micro-physical stuff.

Many psychologists have suggested that we have an innate concept of object and that the concept doesn't really change as we age. Spelke has suggested four principles which characterize our object concept.

- 1) An object is a connected and bounded region of matter that maintains itself when in motion.
- 2) Principle of continuity: Roughly this says that objects do not teleport.
- 3) Principle of contact: Objects don't move unless touched by something that is moving.
- 4) Principle of solidity: One object cannot pass through another object (without damage).

Suppose that something like this is right. Presumably, Aristotle had this concept of object. Now the question is whether he can find anything in the micro-physical description of the world that matches this concept of object. The problem is that the modern view of quantum mechanics has things violating every one of these principles.

Principle 1): From the micro-physics standpoint "matter" is all over the place. The only boundaries we can draw are based on densities. We might be able to come up with a sense of "connected" which involves the forces which make one chunk of matter more dense than another. But nothing in QM is connected in the macro sense of "touching." Finally, matter does not maintain its connectedness while in motion. Electrons leave my body every time I stand up from my couch and touch something metal. In fact the micro-physical elements in every object are constantly being exchanged with elements from other objects as well as being in constant motion within an object.

Principle 2): There is a phenomena in QM known as quantum tunneling. Basically elements at the QM level simply move from one place to

another without existing anywhere in between. At the level of individual QM elements this is happening all the time. Furthermore, there is a finite probability that this could happen to a macro sized object as well. But if our concept of macro sized objects doesn't allow for this then how is our ideal Aristilian reasoner going to figure out that some particular portion of space time currently has an object in it.

I'll skip 3 and 4 as I think the point has been made. We get from everyday objects to particle physics by reductive analysis. Objects are collections of molecules. Molecules are composed of atoms. Atoms are composed of some elementary particles. An ideal reasoner with only Aristotle's concepts and a theory of elementary particle physics won't find objects in a quantum mechanical description of the world. Nothing behaves the way objects do when described by quantum mechanics. In a molecular description of the world we get collections of things that do behave like objects. In an atomic description of the world we get things that behave like molecules. Finally, in a QM description of the world we get things that behave like atoms. But I think you really do need the intervening concepts to get from QM to objects.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Apr 5 16:22:28 1999

Date: Mon, 5 Apr 1999 16:13:33 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik J Larson <erikl@U.ARIZONA.EDU>

Subject: Re: CH and Quantum Mechanics

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Hi everyone,

I've been thinking that Horgan may be too dismissive of the possibility that quantum indeterminacy could spell trouble for the supervenience thesis. He points out that DM and S are "logically independent", and apparently thinks that any quantum indeterminacy in physics would apply only to DH.

I want to attempt to develop a quantum counterexample to CH and the supervenience thesis. Consider a conditional whose antecedent contains a complete specification of microphysical facts for some time slice, and whose consequent contains the set of macrophysical facts entailed by the antecedent. Apply this to the Shrodinger's cat thought experiment, where a photon impinges a set of half-silvered mirrors with probability one-half of deflection or penetration and thereby either triggers or fails to trigger a mechanism which breaks a flask of poison in a sealed container containing a live cat. The idea is that the cat's life status is indeterminate while the photon is superposed over possible states in the deflective mirror apparatus, until someone opens the container, collapsed the superposed "wave-function" and either observes a healthy cat or a dead one. Now, given the interpretation of quantum theory that posits a real indeterminacy (takes a realist interpretation on the wave-function and its collapse), it would seem that there would be one perfectly well-formed macro-fact--the cat being either alive or poisoned to death--that cannot be entailed by the antedent of the CH conditional. The cat example seems contrived (it is) but the breakdown of the CH thesis (S) would seem to follow anytime we want to know macro-facts about the future that have to be entailed from prior microphysical facts that enter into quantum superpositions, generally.

I'm thinking that Horgan dismisses quantum indeterminacy because he

considers only cases of entailment from some set of microphysical facts to some other set, such as the conditional from rates of decay for the particles of a radioactive element to a specification of a particular particle's emission, and other standard examples of quantum indeterminacy.

That being said, the anticipated objection is that some interpretations of quantum theory do not take a realist line on the wave-function collapse. (I don't remember all the details of the various interpretations, so I'll say more generally that some interpretations apparently can accommodate the collapse without conceding fundamental indeterminacy). Still, if it is at least plausible that quantum theory may turn out to be fundamental to a completed physics, then in that conceivable scenario the supervenience thesis of CH will fail to hold, and so the argument against materialism--predicated on the distinction between failed supervenience only in the phenomenal case and not for standard physical macro objects--would seem to be in big trouble. This strikes me as odd, since the cogency of the argument would then hinge on the contingencies of our physical theory qua quantum indeterminacy. It seems strange to say that IF our physical theory turns out such and such, then we have a good argument against materialism, but possibly not. It also seems strange that the anti-materialist argument should be put forth so assertively given our epistemic confusion about QM in general, since it seems that the failure of CH and S is at least possible on some interpretations, and in that case the argument wouldn't be very compelling.

Erik

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 6 15:15:32 1999
Date: Tue, 6 Apr 1999 15:10:59 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: Re: reductive concepts
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Some thoughts on Josh's email. I'm not sure that the features of quantum mechanics Josh notes are necessarily all that problematic for cosmic hermeneutics. Although I do not know all that much about QM, I think that, when we say that an atom is a certain size, we specify a sphere within which there is a fairly good chance of finding the appropriate electrons. Although this certainly allows that there is some amount of fuzziness in the boundaries of objects, this does not necessarily undermine the possible success of cosmic hermeneutics-- given the exceedingly high probability of electrons and so forth occupying particular regions of space, the 'macro' properties of solidity and so forth arise.

I think the root of Josh's worries seems to be the quantum indeterminacy stuff. The Heisenberg Uncertainty Principle asserts that we cannot know

both the location of a particle and its velocity beyond a certain level of accuracy. This arises because, in the process of measuring, we must interact with the particle-- if we measure the velocity of a particle, we alter its position, and so forth. A Laplacian Demon, I presume, is not constrained by this as it simply knows all of the microphysical facts of the universe at a particular instant. I wonder whether this bypasses the problems of indeterminacy with which we are faced when we study quantum phenomena.

Anthony

> Principle 1): From the micro-physics standpoint "matter" is all over the
> place. The only boundaries we can draw are based on densities.
> We might be able to come up with a sense of "connected" which involves
> the forces which make one chunk of matter more dense than another.
> But nothing in QM is connected in the macro sense of "touching."
> Finally, matter does not maintain its connectedness while in motion.
> Electrons leave my body every time I stand up from my couch and touch
> something metal. In fact the micro-physical elements in every object
> are constantly being exchanged with elements from other objects as
> well as being in constant motion within an object.

> Principle 2): There is a phenomena in QM known as quantum tunneling.
> Basically elements at the QM level simply move from one place to
> another without existing anywhere in between. At the level of
> individual QM elements this is happening all the time. Furthermore,
> there is a finite probability that this could happen to a macro sized
> object as well. But if our concept of macro sized objects doesn't
> allow for this then how is our ideal Aristotelian reasoner going to
> figure out that some particular portion of space time currently has an
> object in it.

>
> I'll skip 3 and 4 as I think the point has been made. We get from
> everyday objects to particle physics by reductive analysis. Objects
> are collections of molecules. Molecules are composed of atoms. Atoms
> are composed of some elementary particles. An ideal reasoner with
> only Aristotle's concepts and a theory of elementary particle physics
> won't find objects in a quantum mechanical description of the world.
> Nothing behaves the way objects do when described by quantum
> mechanics. In a molecular description of the world we get collections
> of things that do behave like objects. In an atomic description of
> the world we get things that behave like molecules. Finally, in a QM
> description of the world we get things that behave like atoms. But I
> think you really do need the intervening concepts to get from QM to
> objects.
>

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 6 21:56:29 1999

Date: Tue, 6 Apr 1999 21:53:57 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Re: CH and Quantum Mechanics

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Lots of interesting issues about QM and cosmic hermeneutics. Josh suggests that one cannot get from QM to "there are macro objects" a priori, as the concept of an object involves various principles (connectedness, continuity, contact, solidity) that are false in QM.

I think it's certainly right that QM poses problems here. When

philosophers talk about microphysics in this context they usually assume some sort of classical mechanics. This makes things easier, as one has definite entities with definite position, velocity, mass, etc, right at the bottom level. From there, it's easy to get to macro position, velocity, shape, size, mass density, etc. But from QM, this isn't so easy.

Of course one thing one might do is cheat, and say: in the base, we'll allow not just microphysics but chemistry. Then we'll be allowed facts about definite molecules with definite positions, masses, etc. And from there we can get the macro stuff out. We still have an interesting and substantive thesis, which will still probably capture what's at issue between someone like me (or Jackson, Lewis, etc), and someone like Byrne, Block, or Stalnaker. But still, this leaves open the interesting question of just what follows from microphysics alone, whether chemistry follows from microphysics, etc.

On Josh's principles: of course violations at the micro level alone may not matter too much, as it's at the macro level that we expect to find objects. But serious violations at the macro level will matter. One strategy would be to argue that such violations won't happen, or will happen only rarely. E.g., quantum tunnelling discontinuities are presumably rare at the macro level, so they may not rule out objects there. And maybe (maybe) there is some way to make sense of connectedness using the QM notion of decoherence, etc. So I think there are strategies here.

A more detailed answer depends really on what interpretation of QM one gives. Without such an interpretation, one can't really know what QM is saying about microphysics. With such an interpretation, I think one gets different results re cosmic hermeneutics.

(1) Bohm interpretation (with particles governed by the wave function): Here there isn't too much trouble for cosmic hermeneutics. One has definite particles all the way down, just as on a classical view. So no more problems deriving macro objects here than on a classical view.

(2) Everett interpretation (with a giant superposed wave function with all possibilities at once): Here, one certainly can't derive macro "truths" from the wave function. E.g., the "truth" that the cat is alive or the point is pointing up won't follow from the superposed wave function. But arguably, on the Everett view these aren't really truths. The cat is really a superposition of both alive and dead in reality, and so on. And of course we don't expect indeterminate matters to be settled by cosmic hermeneutics on microphysics, any more than questions of tallness are.

Alternatively, one can say that the sense in which the cat is "really" alive is relative to my own mind -- my mind splits into a component, one of which sees a dead cat and one (this one) an alive cat. If so, the cat status won't follow from microphysics, but it will follow from microphysics plus facts about my mind. What to say here depends on whether mental facts are derivable from physical facts. If they are derivable, then the cat status will follow from physical facts plus an indexical about where *I* am. If they are not derivable, then we need mental (e.g. phenomenal) truths in the inference base. I take the latter option, of course. Basically, on this version of the Everett view, cosmic hermeneutics is radically false, as the apparent "determinacy" in the world stems from the minds of observers and not from the world. But one still has derivability from the microphysical

plus mental combined.

(3) Collapse interpretation (on which there is just a wave function which "collapses" into a relatively definite state every now and then): This is probably the hardest case. The collapse will be in microphysics, so one will be able to know that the pointer or cat wavefunction has collapsed. But there is still a question of how one gets from even a collapsed wavefunction to macro truths -- e.g. from a macro object's wave function to facts about its position -- and whether this is a priori. The position etc of an object or particles will correspond to a certain eigenvalue, but it's not clear in what sense the mapping from microphysics to position is a priori, as it's not obvious in what sense taking the eigenvalue is part of the concept.

Perhaps the easiest thing is to build this matter into the interpretation itself, so that the wavefunction is to be counted as specifying these positions etc via eigenvalues (that's part of what the wavefunction means). If one does that, then arguably one can get from the wave function to positions, velocities, etc, of various entities, and then eventually to macro determinacies, at least in some cases (cases where the wave function has collapsed appropriately, so that things actually have a determinate position, etc).

The alternative is to say that we can't get this from the wave function alone, and that we need to introduce substantive "bridge principles" to get from wave function to macro position, etc. On the previous view, these bridge principles are a priori, but if that's denied, then they are substantive and a posteriori. On such a view cosmic hermeneutics will need to be denied. But arguably that's appropriate. The "thin" nature that we are building into microphysics on this interpretation arguably needs to be supplemented with further laws (these bridge laws) to determine a macro world. So the bridge laws should be taken as a further part of the theory. I have heard some people (e.g. Shelly Goldstein at Rutgers) argue for this explicitly, and some others seem to be implicitly committed to such a view.

Personally, I prefer the former view, but I think it is a difficult and somewhat open question. The ontology of these collapse interpretations has not been worked out as well as it might be, I think.

Overall: Clearly the answer depends on the interpretation. But I think we find a pattern: in the cases where the microphysics gives us a **complete** theory of the fundamental laws and such (e.g. Bohm, some collapse interpretations), cosmic hermeneutics is possible. In cases where microphysics is an **incomplete** theory, and where further fundamental laws are needed (e.g. some collapse and Everett interpretations), cosmic hermeneutics is impossible. But in each case, the macro truths (insofar as there are such truths) follow a priori from the fundamentals, whatever those are.

Presumably, given an interpretation of QM, with all the relevant concepts, and with all the relevant facts, even Aristotle would be able to figure out the truths about macro objects (insofar as there are such truths) from the relevant facts. In the Bohm case it's easy; in the Everett case he may need facts about minds; in the collapse case he may or may not need further principles, depending on just what we give him as "built into" the microphysical concepts. But in each case some sort of macro picture should emerge, insofar as there is such a picture (of course, in some version of the Everett

interpretation, there simply is no macro picture in the world as opposed to in the mind).

Erik L. considers a case involving Schrodinger's cat, where an observation triggers a collapse in the cat's wave function at time t , so it goes from a superposition of alive and dead to being alive (let's say). So we're assuming a collapse interpretation here. Erik suggests that the facts that the cat is alive isn't entailed by the microphysics here.

I think it's right that the cat's status isn't entailed by microphysics before time t . But that's not a problem for cosmic hermeneutics; rather, it's a problem for "cosmic number-crunching". In this case, Laplace's number-crunching demon can't even predict the future evolution of the wave function from the previous state. But the hermeneutic demon is given **all** the facts about the wavefunction across space and time, so this isn't a problem for him.

Given all the microphysical facts (including facts after t), presumably examination of the post- t wave function will reveal that the cat is alive rather than dead. One will find particles with definite locations constituting a cat walking around, etc, with a beating heart, etc, rather than a cat lying down with no heartbeat, etc. At least, that's so modulo the issues about interpretation of collapse formalisms discussed above. But any such problems won't stem merely from indeterminate evolution between past and future states of the wave function, which the interpreting demon can take for granted.

Erik suggests that given the possibility that CH fails for QM, one should be careful putting forward anti-materialist arguments given the failure of CH in the phenomenal case. I do agree that QM should make us cautious here. But I'd like to think that on a careful examination of QM, the interpretations on which CH fails are precisely those in which the relevant microphysics is an incomplete theory. So one still has an inference from failure of CH to expansion of ontology.

Anthony notes another way of resisting the conclusion that QM falsifies CH. If in QM, particles have "fuzzy" location, that may be enough to give enough macro determinacy for our purposes. I think that's more or less right, though the details of this fuzziness depend a lot on the interpretation of QM. On the Bohm interpretation, the particles have definite interpretations and the wave function corresponds to our epistemic probabilities of finding them in certain definite places, so one has probabilistic access to the truth of the sort Anthony suggests. Of course in this case the demon will strictly speaking have access to the positions of the underlying particles, not just to the wave function, so he won't have problems. As Anthony notes, the demon is not constrained by our limitations of measurement -- he knows all the determinate facts!

On other interpretations, things are a bit more complicated, as there really aren't particles with determinate positions in the general case. Here it's not just a problem of access to the facts, but of there not being facts. So one can't really even talk of "the probability that the particle is in position X ". The wave function gives the full state of the particle, and that's that. But still, some of the time, the wave function will specify a definite position (as eigenvalue of position operator), which will help. And other times, it will specify a determinate range of positions (as a range of

eigenvalues), which will presumably correspond to the sort of "fuzzy" location Anthony mentions. And maybe such micro fuzziness could add up to a sort of macro determinacy.

There is a worry that some such fuzziness can be amplified at the macro level rather than cancelled out. That gets quite tricky, and again depends on the QM interpretation. On collapse interpretations, it will be the case that at least sometimes, the amplified fuzziness will be collapsed out, so in those cases we'll have a relatively determinate state. On Everett interpretations, the fuzziness may amplify indefinitely and forever (!), so there really won't be any determinacy to be found in the physical state alone; any determinacy will then have to be found in the mind.

Anyway, plenty of issues here.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 8 01:53:46 1999

X-Priority: 3

X-MSMail-Priority: Normal

X-MimeOLE: Produced By Microsoft MimeOLE V4.72.3110.3

Date: Thu, 8 Apr 1999 01:48:47 -0700

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Brad Thompson <bradt@U.ARIZONA.EDU>

Subject: Does the CH-thesis have metaphysical significance?

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

In looking back over the Horgan and Byrne readings on cosmic hermeneutics, I am somewhat puzzled about the significance of the CH-thesis. It seems that Horgan is interested in the possibility of CH because it provides a way of capturing the physicalist claim that all facts are necessitated by the microphysical facts. Now, there is no *immediately* obvious contradiction in claiming that all the facts are necessitated (this is a metaphysical notion) by the microphysical facts and yet for a demon (an ideal reasoner) who knows all these microphysical facts to be unable to deductively infer from this knowledge all true propositions. I say there is no obvious contradiction in holding these two views because the former is about metaphysics and the latter is about epistemology.

Nonetheless, this physicalism minus CH seems to create a gap between reality and our knowledge of reality (or at least its intelligibility) which is extremely unpalatable. The problem is that those who deny CH have to claim that we arrive at micro to macrophysical identity statements (I hope this is an ok way of characterizing "water = H₂O") in some other way than the analysis approach suggested by Chalmers, Jackson, and even Loar. The analysis approach renders the identity intelligible. But I suspect that any other way of arriving at such identities falls short of intelligibility.

This came out of discussion of the Byrne article a couple weeks ago--the denial of an a priori entailment of the macro from the micro leads to a new skeptical hypothesis, such as the "problem of other chairs"! If I know all the microphysical facts of the thing I'm currently sitting on, I can still wonder "Is it a chair?" Here is a more specific way of bringing out this point. Suppose, as I think Block, Stalnaker, and Byrne would probably say, we arrive at identities through the observation of correlations. Chemists observe a correlation between "being a water sample" and "being an H₂O sample" and conclude on grounds of parsimony perhaps that "water = H₂O". But if our reasons for believing in the identity claim is merely that they are correlated, then we can certainly raise the worry that the property of

this stuff in virtue of which it falls within the (secondary) intension of "water" is a different property from the property in virtue of which it falls within the intension of "H2O".

I believe that Block and Stalnaker claim that identities don't need explanation. But they also seem committed to the view that identities can't be understood--they just have to be taken as brute. Physicalism minus CH seems to lead to the view that there are strong necessities everywhere, and that our hermeneutic demon would have to know the microphysical facts PLUS these strong necessities in order to deduce all the facts. This really seems outside the spirit of physicalism.

Brad

Brad Thompson
Department of Philosophy
University of Arizona
Tucson, AZ 85721-0027

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 8 12:41:02 1999

Date: Thu, 8 Apr 1999 12:38:58 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik J Larson <erikl@U.ARIZONA.EDU>

Subject: Re: Does the CH-thesis have metaphysical significance?

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

Brad,

I'm going to pick on just a specific part of your message here; namely, your point that micro to macro identity statements might be had by taking note of correlations between properties. I would argue that mere correlations cannot possibly constitute sufficient grounds for identity. For example, there is a high correlation between the occurrence of a storm and a rapid fall in barometric pressure, but claiming an identity here would be absurd. However we arrive at identities, it can't be from just observed correlations, even if this constitutes an initial heuristic for focusing attention on certain properties over others.

Erik

On Thu, 8 Apr 1999, Brad Thompson wrote:

> In looking back over the Horgan and Byrne readings on cosmic hermeneutics, I
> am somewhat puzzled about the significance of the CH-thesis. It seems that
> Horgan is interested in the possibility of CH because it provides a way of
> capturing the physicalist claim that all facts are necessitated by the
> microphysical facts. Now, there is no *immediately* obvious contradiction
> in claiming that all the facts are necessitated (this is a metaphysical
> notion) by the microphysical facts and yet for a demon (an ideal reasoner)
> who knows all these microphysical facts to be unable to deductively infer
> from this knowledge all true propositions. I say there is no obvious
> contradiction in holding these two views because the former is about
> metaphysics and the latter is about epistemology.
>

> Nonetheless, this physicalism minus CH seems to create a gap between reality
> and our knowledge of reality (or at least its intelligibility) which is
> extremely unpalatable. The problem is that those who deny CH have to claim
> that we arrive at micro to macrophysical identity statements (I hope this is

> an ok way of characterizing "water = H2O") in some other way than the
> analysis approach suggested by Chalmers, Jackson, and even Loar. The
> analysis approach renders the identity intelligible. But I suspect that any
> other way of arriving at such identities falls short of intelligibility.
>
> This came out of discussion of the Byrne article a couple weeks ago--the
> denial of an a priori entailment of the macro from the micro leads to a new
> skeptical hypothesis, such as the "problem of other chairs"! If I know all
> the microphysical facts of the thing I'm currently sitting on, I can still
> wonder "Is it a chair?" Here is a more specific way of bringing out this
> point. Suppose, as I think Block, Stalnaker, and Byrne would probably say,
> we arrive at identities through the observation of correlations. Chemists
> observe a correlation between "being a water sample" and "being an H2O
> sample" and conclude on grounds of parsimony perhaps that "water = H2O".
> But if our reasons for believing in the identity claim is merely that they
> are correlated, then we can certainly raise the worry that the property of
> this stuff in virtue of which it falls within the (secondary) intension of
> "water" is a different property from the property in virtue of which it
> falls within the intension of "H2O".
>
> I believe that Block and Stalnaker claim that identities don't need
> explanation. But they also seem committed to the view that identities can't
> be understood--they just have to be taken as brute. Physicalism minus CH
> seems to lead to the view that there are strong necessities everywhere, and
> that our hermeneutic demon would have to know the microphysical facts PLUS
> these strong necessities in order to deduce all the facts. This really
> seems outside the spirit of physicalism.
>
> Brad
>
>
> _____
> Brad Thompson
> Department of Philosophy
> University of Arizona
> Tucson, AZ 85721-0027
>

"What our grammarian does is simple enough. He frames his formal reconstruction of K along the grammatically simplest lines he can, compatibly with inclusion of H, plausibility of the predicted inclusion of I, plausibility of the hypothesis of inclusion of J, and plausibility, further, of the exclusion of all sequences which ever actually do bring bizarreness reactions." -- W.V.O. Quine

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 8 16:37:33 1999
Date: Thu, 8 Apr 1999 16:35:08 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Brad J Thompson <bradt@U.ARIZONA.EDU>
Subject: Re: Does the CH-thesis have metaphysical significance?
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I agree with what Erik says below, concerning the shortcomings of arriving at identities via observed correlations. That really was my point. Correlations seemed to be the best that Block, Stalnaker, and Byrne can

do to establish identities, and that seems unsatisfactory. So how else can we arrive at identities if not via a priori conceptual analysis which connects the two terms?

Brad

On Thu, 8 Apr 1999, Erik J Larson wrote:

> Brad,
>
> I'm going to pick on just a specific part of your message here; namely,
> your point that micro to macro identity statements might be had by taking
> note of correlations between properties. I would argue that mere
> correlations cannot possibly constitute sufficient grounds for identity.
> For example, there is a high correlation between the occurrence of a storm
> and a rapid fall in barometric pressure, but claiming an identity here
> would be absurd. However we arrive at identities, it can't be from just
> observed correlations, even if this constitutes an initial heuristic for
> focusing attention on certain properties over others.
>
> Erik
>
>
> On Thu, 8 Apr 1999, Brad Thompson wrote:
>
> > In looking back over the Horgan and Byrne readings on cosmic hermeneutics, I
> > am somewhat puzzled about the significance of the CH-thesis. It seems that
> > Horgan is interested in the possibility of CH because it provides a way of
> > capturing the physicalist claim that all facts are necessitated by the
> > microphysical facts. Now, there is no *immediately* obvious contradiction
> > in claiming that all the facts are necessitated (this is a metaphysical
> > notion) by the microphysical facts and yet for a demon (an ideal reasoner)
> > who knows all these microphysical facts to be unable to deductively infer
> > from this knowledge all true propositions. I say there is no obvious
> > contradiction in holding these two views because the former is about
> > metaphysics and the latter is about epistemology.
> >
> > Nonetheless, this physicalism minus CH seems to create a gap between reality
> > and our knowledge of reality (or at least its intelligibility) which is
> > extremely unpalatable. The problem is that those who deny CH have to claim
> > that we arrive at micro to macrophysical identity statements (I hope this is
> > an ok way of characterizing "water = H₂O") in some other way than the
> > analysis approach suggested by Chalmers, Jackson, and even Loar. The
> > analysis approach renders the identity intelligible. But I suspect that any
> > other way of arriving at such identities falls short of intelligibility.
> >
> > This came out of discussion of the Byrne article a couple weeks ago--the
> > denial of an a priori entailment of the macro from the micro leads to a new
> > skeptical hypothesis, such as the "problem of other chairs"! If I know all
> > the microphysical facts of the thing I'm currently sitting on, I can still
> > wonder "Is it a chair?" Here is a more specific way of bringing out this
> > point. Suppose, as I think Block, Stalnaker, and Byrne would probably say,
> > we arrive at identities through the observation of correlations. Chemists
> > observe a correlation between "being a water sample" and "being an H₂O
> > sample" and conclude on grounds of parsimony perhaps that "water = H₂O".
> > But if our reasons for believing in the identity claim is merely that they
> > are correlated, then we can certainly raise the worry that the property of
> > this stuff in virtue of which it falls within the (secondary) intension of
> > "water" is a different property from the property in virtue of which it

> > falls within the intension of "H2O".
> >
> > I believe that Block and Stalnaker claim that identities don't need
> > explanation. But they also seem committed to the view that identities can't
> > be understood--they just have to be taken as brute. Physicalism minus CH
> > seems to lead to the view that there are strong necessities everywhere, and
> > that our hermeneutic demon would have to know the microphysical facts PLUS
> > these strong necessities in order to deduce all the facts. This really
> > seems outside the spirit of physicalism.
> >
> > Brad
> >
> > _____
> > Brad Thompson
> > Department of Philosophy
> > University of Arizona
> > Tucson, AZ 85721-0027
> >
> >
> >
> "What our grammarian does is simple enough. He frames his formal
> reconstruction of K along the grammatically simplest lines he can,
> compatibly with inclusion of H, plausibility of the predicted inclusion
> of I, plausibility of the hypothesis of inclusion of J, and plausibility,
> further, of the exclusion of all sequences which ever actually do bring
> bizarreness reactions." -- W.V.O. Quine
>
> -----
> Erik J Larson
> erikl@U.Arizona.EDU
>

Brad J Thompson
bradt@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Sat Apr 10 16:35:29 1999
Date: Sat, 10 Apr 1999 16:35:17 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Does the CH-thesis have metaphysical significance?
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I'm obviously in sympathy with what Brad wrote re cosmic hermeneutics. One of the problems with denying CH, I'd argue, is that it creates apparent epistemological problems where we had none before. If there are two different epistemic possibilities compatible with all the physical (and phenomenal) facts, it's not clear how we are to choose between them. At least we have a substantive skeptical problem. But on the face of it there is no such problem for chairs, tables, etc. (Of course there is such a problem relative to the phenomenal facts alone, but not relative to the physical and phenomenal facts.)

There is an interesting question about how one gets to micro-macro identity without CH. One thing Block/Stalnaker et al might say is that one gets such an identity when one has a reductive explanation, but deny that reductive explanation is to be cashed out in terms of a priori entailment. Another thing would be to deny that reductive explanation is required at all, as that's an epistemic matter where

identity is ontological. I would tend to agree with Brad that either way, we are left with "brute" identities that are very difficult to integrate with our view of the world and that border on being unintelligible.

Block/Stalnaker et al would argue that these identities are perfectly intelligible once we get out of our a priori straightjacket. Combinations of empirical evidence plus considerations of simplicity can make such identities very plausible even where there is no transparent epistemic connection from micro to macro. One might say that the identity is the "best explanation" of the empirical evidence, including correlation. And they would argue that it is OK for an identity to be "brute" in this way, as identities don't need to be explained. They use "Samuel Clemens = Mark Twain" as an example here -- how would one explain that? But they'd say that nevertheless, such identities are perfectly intelligible.

Now, I'd argue that for standard identities, some sort of epistemic entailment from a more fundamental base still holds (e.g. one could deduce that Mark Twain is Samuel Clemens from all the physical plus phenomenal facts), and this is what renders the identity there intelligible and "non-brute". Where the identity needed in the mind-brain case would seem quite "brute". But clearly B&S don't agree with either of these points.

I don't think B&S need say that every correlation implies an underlying identity. They could deal with some cases (e.g. the ones Erik L. raises) by putting some conditions on them. E.g., one would at least require the correlated events to occur at the same point in space and time. And they would argue that it can't be the case that there is an obviously causal process mediating the relation between the two (as with storm and pressure). And so on. So maybe they could motivate identities in the water-H₂O cases and mind-brain cases without being committed to every correlation implying identity.

N.B. One note: Physicalism minus CH need not quite require strong necessities as defined (these involve positive conceivability without possibility), but they will at least lead to inscrutable truths (which come down to negative conceivability without possibility). I don't think inscrutable truths are quite as problematic as strong necessities, but they still have lots of problems, and in particular seem unsuited to play much of a role in epistemic matters such as as explanation.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Sat Apr 3 12:55:28 1999
Date: Sat, 3 Apr 1999 13:53:59 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: dogs
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

In Dave's response to Angela re the capabilities of the "ideal conceiver", he suggests that, "...Arguably, I can conceive roughly of what it's like visually to be a dog, even though I'm not a dog." This seems to be accurate. Whereas we have rod and cone structures in our eyes, dogs only have rod structures (I think..). Given that we have good reasons for thinking that cones are responsible for our experience of color, we think that dogs see in black and white. It seems quite straightforward to

imagine the visual experiences of a dog-- they are probably somewhat like the experience of seeing a black and white movie.

It seems to me that it is somewhat more difficult to conceive of a dogs olfactory experiences. Avalanche rescue dogs are capable of quickly detecting the odor of people buried beneath considerable amounts of very hard snow (3 meters or more, sometimes). They are further able to distinguish the smell of the buried victim from the smell of the person handling them and other rescuers on the scene. I was once buried in a snowcave some six feet below the surface, and the rescue dog found me and had dug down to me in about 2 minutes.

I don't have the faintest idea what it would be like to be able to make the kinds of olfactory discriminations that dogs are capable of making. It does not seem to be me to be particularly likely that even an ideal conceiver would be able to do this. I seems fairly straightforward to imagine a what a dog's visula experiences are like because we assume that their visual experiences are somehow less rich than our own. But, given that dogs have olfactory capabilities far superior to ours, itjust doesn't seem plausible to suggest that any non-canine being, however idealized and with whatever instrumentation you please, could conceive of what a dogs olfactory experiences are like.

I'm not sure if all of this amounts to anything (except praise for the virtues of dogs).

Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Wed Mar 31 00:29:40 1999
Date: Wed, 31 Mar 1999 00:29:04 -0800
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Next week
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

For next week's meeting, we will focus on possible worlds. The readings should be in the purple folder by sometime Wednesday. They are

- Lewis, Possible worlds
- Adams, Theories of actuality
- Stalnaker, Possible worlds.
- Lewis, Selections from "On the Plurality of Worlds".

These present three different ways of understanding the nature of possible worlds. Lewis argues that possible worlds exist in just the same way that the actual world does (modal realism, or concrete modal realism). Adams and Stalnaker argue that possible worlds should be understood as some sort of abstract objects, or as "ersatz" worlds constructed via a linguistic or propositional construction. (The last part of Adams on "world-stories" is particularly relevant.) The Lewis book selections defend modal realism in more detail, and give a critical discussion of some "ersatz" proposals.

If you can, it might also be good to look at the Lewis book as a whole. This goes exhaustively into all sorts of issues here -- why modality is important (chapter 1), objections to his modal realism (chapter 2), and pros and cons of the various "ersatz" constructions (chapter 3). It's also a wonderful clear book to read. It's out of print, unfortunately, and too long to reproduce wholly here, so I will copy just parts of chapters 2 and 3. The chapter 3 material (which in essence can be seen as a response to the Adams and Stalnaker proposals) is particularly relevant for our discussion.

We'll discuss those general issues in the first half of next week's meeting. In the second half we'll try to apply the issues in the general context of conceivable worlds, two-dimensionalism, etc. In particular, we will see if we can justify the claim that the space of ideally conceivable worlds exists in at least the sense that standard possible worlds exists (e.g. an "ersatz" sense). We'll talk about why such worlds might be needed; and we will see whether we can go through an "ersatz" construction so that this space of worlds makes coherent sense. This will correspond to what I say extremely briefly in Mind and Modality 2.9, (i) and (ii). It would help to look at that, and also at what I say re the need for conceivable worlds in 3.2 of "Materialism and the Metaphysics of Modality". If you can, think about how one might go about making up an ersatz construction (perhaps starting from concepts and the notion of apriority, and perhaps one or two other primitives) that will deliver the right results.

In the meantime, on the mailing list this week it would be good to see a detailed discussion of some of the issues re scrutability and cosmic hermeneutics. E.g. Horgan, Byrne, Block and Stalnaker, my discussion, and any general thoughts you might have. We didn't get a chance to go over Block and Stalnaker exhaustively in class, so it might be particularly useful to hear any comments on the details of their paper.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 6 23:24:05 1999
Date: Tue, 6 Apr 1999 23:22:47 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik A Herman <erikh@U.ARIZONA.EDU>
Subject: possible worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R0

I have a question about possible worlds:

A possible world is the set of stuff that is causally connected in space and time. How do you treat the effect of "consideration" of a possible world on the actual world. For example, this actual world is now different than it would have been had we not considered a possible world where there are talking donkeys. How is this reconciled?

Erik H.

From owner-modality@LISTSERV.ARIZONA.EDU Sat Apr 10 16:38:04 1999
Date: Sat, 10 Apr 1999 16:37:55 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: possible worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R0

On Erik H.'s point re possible worlds: I think it's clearly true that a world where I'm thinking of talking donkeys is different from a world where I'm not. More generally, a world W1 where I'm thinking of world V1 is different from a world W2 in which I'm thinking of world V2. But this doesn't imply that there is causal interaction between V1 and W1. Rather, V1 and W1 have been the way they are all along!

Still, there is a deep question about just how one can think about something that one can't causally interact with. This bothers quite a few people thinking about the epistemology of possible worlds. One can at least say in reply that the problem here isn't obviously worse than the problem in mathematics, where one can apparently think about the number 2 without causally interacting with it. Someone who believes that possible worlds are "abstract objects", like mathematical objects, might well hope that the two problems will have the same sort of solution (whatever that is). Lots of solutions have been offered, but I don't think any single solution has widespread acceptance.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 13 22:34:42 1999
x-sender: agillies@pop.u.arizona.edu
Date: Tue, 13 Apr 1999 13:15:48 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: truth conditions and conceivable worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R0

All,

We ended last time by sketching the truth-conditions for modal sentences using a "conceivable worlds" semantics. Maybe I missed something. I thought the semantics should go something like the following:

(*) Nec P is true₁ iff: for any conceivable world w, P is true₁ at w.

(**) Nec P is true₂ iff: for any conceivable world w, P is true₂ at w.

But (**) doesn't at all correspond to Dave's definition (if I recall). I would have thought the above was the natural way to do things since it recursively calls truth₁ or truth₂ for non-modal sentences, which we already have worked out (the general sketch of the 2D framework does that). The problem is with (**). The semantics is supposed to hold out the possibility that the class of conceivable worlds is not co-extensive with the class of metaphysically possible worlds. But (**) makes appeal to conceivable worlds where P is true₂, and truth₂ is what some might call "metaphysical truth" (type-B materialists, e.g.). Combining that with (**), though, we get that there are no truths P such that Nec P is true₂. And this is awkward for conceivable world semantics. In fact, it might give someone reason to doubt that there is anything useful (apart from defeating materialism) in the semantics. On analogy with possible worlds semantics, conceivable worlds semantics needs the claim to usefulness. Fixing (**) might be problematic too: (**) is the natural way of giving modal semantics, and changing only (**) without substantively changing (*) looks ad hoc.

This points to another confusion I've been entertaining on 2D semantics. It's probably a little off topic, so we'll get to it in a week or two. But the main idea is to what extent the 2D semantics for terms meshes with the 2D semantics for sentences. In particular, in figuring the truth₁ of a sentence S, it looks like every term in S must be mapped to its referent₁. Likewise for truth₂. Are there sentences S for which the natural truth₁ conditions might depend on some term in S being mapped to its referent₂?

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 13 23:44:23 1999
Date: Tue, 13 Apr 1999 23:44:09 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: truth conditions and conceivable worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Thony wrote (it came through eventually):

>We ended last time by sketching the truth-conditions for modal sentences
>using a "conceivable worlds" semantics. Maybe I missed something. I
>thought the semantics should go something like the following:

>
>(*) Nec P is true₁ iff: for any conceivable world w, P is true₁ at w.
>

>(**) Nec P is true₂ iff: for any conceivable world w, P is true₂ at w.
>

>But (**) doesn't at all correspond to Dave's definition (if I recall).

I don't remember exactly what I said last week. But looking at what

you have here: it's not a million miles from what I would say, but I would change some details. In particular, I wouldn't put the left-hand sides as "'nec P' is true₁" and "'nec P' is true₂", as the two right hand sides can have different actual truth-values for many P, but the left-hand sides cannot, at least at the actual world. (If a statement is true₁ at the actual world it is true₂ there, and vice versa.) So it's better to write the left-hand sides as "P is 1-necessary" and "P is 2-necessary", or "nec₁ P" and "nec₂ P", or some such.

>I would have thought the above was the natural way to do things since it
>recursively calls truth₁ or truth₂ for non-modal sentences, which we
>already have worked out (the general sketch of the 2D framework does
>that).

For the reasons above, I think it's best not to think of 1-necessity and 2-necessity as the primary and secondary intensions of a single concept of "necessity". Rather, it's that the concept of "necessity" is ambiguous, and we need to disambiguate two distinct concepts. Probably these concepts are best regarded as not themselves having a deep 2-D structure -- it's not as if the notion of "necessity" (or "1-necessity" or "2-necessity") picks out different things depending on how the actual world turns out. At least, nothing in the 2-D considerations alone suggest that conclusion, though one might conceivably come up with other arguments for it.

>The problem is with (**). The semantics is supposed to hold out
>the possibility that the class of conceivable worlds is not co-extensive
>with the class of metaphysically possible worlds. But (**) makes appeal
>to conceivable worlds where P is true₂, and truth₂ is what some might
>call "metaphysical truth" (type-B materialists, e.g.). Combining that
>with (**), though, we get that there are no truths P such that Nec P is
>>true₂. And this is awkward for conceivable world semantics.

I don't quite follow your claim at the end. Even with the semantics you suggest above, won't "2+2=4" and "water is H₂O" be examples of truths P such that Nec P is true₂ (as P will be true₂ in all conceivable worlds)?

But anyway, I think one has to be careful about equating truth₂ with "metaphysical truth". Actually, I'm not sure what "metaphysical truth" is, but I imagine the above comes to much the same thing as equating 2-necessity with "metaphysical necessity", so I'll consider that claim instead. It's true that on the 2-D picture, and on the modal rationalist view, 2-necessity (as defined here) and metaphysical necessity more or less coincide (at least on one plausible way of reading the latter). But on a view with strong necessities, the two may come apart, as such a view will have fewer metaphysically possible worlds than conceivable worlds. For example, it might just be that "the gravitational constant = XXX" is metaphysically necessary but not 2-necessary as defined here.

On views with strong necessities, I suppose P will be metaphysically necessary is P is true₂ in all metaphysically possible worlds. Given that the spaces of conceivable and metaphysically possible worlds come apart on such a view, metaphysical necessity and 2-necessity (as defined above) will come apart similarly.

>In fact,
>it might give someone reason to doubt that there is anything useful
>(apart from defeating materialism) in the semantics. On analogy with
>possible worlds semantics, conceivable worlds semantics needs the claim

>to usefullness. Fixing (**) might be problematic too: (**) is the
>natural way of giving modal semantics, and changing only (**) without
>substantively changing (*) looks ad hoc.

I don't quite follow this, but maybe you can elaborate. Again, I prefer to simply disambiguate the concept of necessity, rather than give it a 2-D modal structure. I think one can make the case for such ambiguity on independent grounds in any case, as e.g. we'll do this week in the material on subjunctives and indicatives.

>This points to another confusion I've been entertaining on 2D semantics.
>It's probably a little off topic, so we'll get to it in a week or two.
>But the main idea is to what extent the 2D semantics for terms meshes
>with the 2D semantics for sentences. In particular, in figuring the
>truth_1 of a sentence S, it looks like every term in S must be mapped to
>its referent_1. Likewise for truth_2. Are there sentences S for which
>the natural truth_1 conditions might depend on some term in S being
>mapped to its referent_2?

Hmm, interesting. Again, a statement is true_1 at the actual world iff it is true_2 there. And a term's (actual) referent_1 will always be its (actual) referent_2 -- both are just the term's actual referent! It's only in non-actual possible worlds that these things come apart. But I suppose your suggestion might come to the possibility that to evaluate the truth_1 of S at some non-actual possible world, we need to consider the referent_2 of some term in S there, which would then depend on a posteriori facts about the actual world.

The short answer, I think, is that this can't happen. Truth_1 is defined so that it depends only on facts about the world in question (plus a priori analysis). Some statements, e.g. "The actual president is Bill Clinton" might make reference to the actual world, but they can still be evaluated in a single-world way. E.g. to evaluate the truth_1 of the statement above at a world where George Bush is president, we consider that world as actual, and determine that if that world is actual, the statement is false. That's to say, one never needs to import the *actual* referent of "the actual president" in evaluating the truth_1 of this statement at world W. One only ever imports what the referent would be *if* W were actual.

Of course one could try to come up with sentences that don't work this way, but I think that more or less by virtue of the way truth_1 is defined, the best we'll see is a pattern akin to the above. But I'm interested to see attempts, in any case!

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Mon Apr 12 19:55:21 1999
Date: Mon, 12 Apr 1999 19:53:52 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Rachael J Parkinson <rachaelp@U.ARIZONA.EDU>
Subject: Re: possible worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I have two questions- In seminar we were talking about other possible worlds which have the same history or are on the same time as our world (for example, consider a world where all the history up to this point is exactly identical with ours, only this seminar never took place.) It seems that this world is spatio-temporally related to ours? If so, should we

consider it as another possible world or part of this world?

One could even make the argument that this no-seminar world is causally related to ours if, from considering the other possible world, my actions are affected. (I imagine how unbarable life would be without this seminar and so decide that I should take it.) Likewise, we could imagine this type of causal interaction with respect to Erik H's talking donkey world.

Finally, in response to Erik H, David suggested that the problem Erik poses can be compared to problems in mathematics... we can contemplate the number 2 without causally interacting with it. But, as Dave points out, this seems a more viable solution if we consider possible worlds as abstract entities, like numbers. Does the problem of how one can think about something one can't causally interact with remain if we consider actual worlds as literal, the way Lewis does?

-Rachael

On Sat, 10 Apr 1999, David Chalmers wrote:

> On Erik H.'s point re possible worlds: I think it's clearly true that
> a world where I'm thinking of talking donkeys is different from a
> world where I'm not. More generally, a world W1 where I'm thinking of
> world V1 is different from a world W2 in which I'm thinking of world
> V2. But this doesn't imply that there is causal interaction between
> V1 and W1. Rather, V1 and W1 have been the way they are all along!
>
> Still, there is a deep question about just how one can think about
> something that one can't causally interact with. This bothers quite a
> few people thinking about the epistemology of possible worlds. One
> can at least say in reply that the problem here isn't obviously worse
> than the problem in mathematics, where one can apparently think about
> the number 2 without causally interacting with it. Someone who
> believes that possible worlds are "abstract objects", like
> mathematical objects, might well hope that the two problems will have
> the same sort of solution (whatever that is). Lots of solutions have
> been offered, but I don't think any single solution has widespread
> acceptance.
>
> --Dave.
>

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 13 01:20:52 1999

X-Priority: 3

X-MSMail-Priority: Normal

X-MimeOLE: Produced By Microsoft MimeOLE V4.72.3110.3

Date: Tue, 13 Apr 1999 01:18:33 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Brad Thompson <bradt@U.ARIZONA.EDU>

Subject: Re: possible worlds

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

In response to one of Rachael's questions:

>I have two questions- In seminar we were talking about other possible
>worlds which have the same history or are on the same time as our world
>(for example, consider a world where all the history up to this point is
>exactly identical with ours, only this seminar never took place.) It seems
>that this world is spatio-temporally related to ours? If so, should we

>consider it as another possible world or part of this world?

I don't think that we should say that this other world is spatio-temporally related to ours. For one, that other world does not appear to be spatially related to this one. Further, I think it's best to restrict temporal relatedness to within-world events. I remember there was debate about this in seminar. But I don't find any problems with this restriction, nor does it really make sense to me to say that two events that are in different worlds occur at the same time in some absolute sense. Rather, when we describe two worlds that have the same history up until a particular time (within each world), events across worlds that we want to describe as "happening at the same time" ought to be described rather as occupying symmetrical locations in their respective world-histories.

Brad

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 13 18:54:35 1999

X-Sender: agillies@pop.u.arizona.edu (Unverified)

Date: Tue, 13 Apr 1999 18:36:23 -0700

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: "Anthony S. Gillies" <agillies@U.ARIZONA.EDU>

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: RO

--===== _-1288068678==_ma=====

Content-Type: text/plain; charset="us-ascii"

All, this is Josh's posting that he was unable to send:

Some thoughts on possible worlds.

Lewis distinguishes between two kinds of parsimony. A theory is qualitatively parsimonious if it posits fewer kinds of entities. A theory is quantitatively parsimonious if it posits fewer tokens of the kinds it posits. His claim is that people are interested in qualitative not quantitative parsimony.

Lewis thinks that his realism about possible worlds is qualitatively, but not quantitatively parsimonious. "You believe in our actual world already. I ask you to believe in more things of that kind, not in things of some new kind." (p185) I'm curious whether that claim is really true. I don't know if this makes sense but see what you think.

There seem to be two kinds of things that a Lewisian view posits that belief in only one actual world doesn't posit. First is logical space. On a one world view logical space is perhaps a concept or an abstract object or whatever. But whatever kind of thing it is, it is something we already have. It isn't clear this is true for Lewis. Logical space seems to at least be a set of super-universal laws that place constraints on how possible worlds can be. Consider the law of non-contradiction. If possible worlds are concepts or abstract objects then non-contradiction is just like other things which govern concepts or abstract objects in our world. But for Lewis this is a law external to our world or any other world. It has to be a different *kind* of law. If purely logical laws bug you then consider Lewis's claim that all worlds have some kind of space time. Certainly, this isn't a logical truth. So if true it has to be a law which governs how possible worlds can be. This is a different kind of

thing than the laws within our world.

The second kind of thing that a Lewisian view posits that a one world view doesn't is entities that we are not spatio-temporally or causally connected to in any way. We have some kind of spatio-temporal connection to everything that exists in our world, including the world itself. Perhaps we have some logical or conceptual connections to other worlds and their contents. But there will be entities which exist yet I have no spatio-temporal connection to. This is a different kind of entity.

I'm not sure how compelling this is but Lewis does seem to be a bit less parsimonious than he thinks.

--===== _-1288068678==_ma=====

Content-Type: text/enriched; charset="us-ascii"

All, this is Josh's posting that he was unable to send:

<fontfamily><param>Helvetica</param><smaller>Some thoughts on possible worlds.

</smaller>Lewis distinguishes between two kinds of parsimony. A theory is

qualitatively parsimonious if it posits fewer kinds of entities. A theory is quantitatively parsimonious if it posits fewer tokens of the kinds it posits. His claim is that people are interested in qualitative not quantitative parsimony.

Lewis thinks that his realism about possible worlds is qualitatively, but not quantitatively parsimonious. "You believe in our actual world already. I ask you to believe in more things of that kind, not in things of some new kind." (p185) I'm curious whether that claim is really true. I don't know if this makes sense but see what you think.

There seem to be two kinds of things that a Lewisian view posits that belief in only one actual world doesn't posit. First is logical space. On a one world view logical space is perhaps a concept or an abstract object or whatever. But whatever kind of thing it is, it is something we already have. It isn't clear this is true for Lewis.

Logical space seems to at least be a set of super-universal laws that place constraints on how possible worlds can be. Consider the law of non-contradiction. If possible worlds are concepts or abstract

objects then non-contradiction is just like other things which govern concepts or abstract objects in our world. But for Lewis this is a law external to our world or any other world. It has to be a different *kind* of law. If purely logical laws bug you then consider Lewis'es claim that all worlds have some kind of space time. Certainly, this isn't a logical truth. So if true it has to be a law which governs how possible worlds can be. This is a different kind of thing than the laws within our world.

The second kind of thing that a Lewisian view posits that a one world view doesn't is entities that we are not spatio-temporally or causally connected to in any way. We have some kind of spatio-temporal connection to everything that exists in our world, including the world itself. Perhaps we have some logical or conceptual connections to other worlds and their contents. But there will be entities which exist yet I have no spatio-temporal connection to. This is a different kind of entity.

I'm not sure how compelling this is but Lewis does seem to be a bit less parsimonious than he thinks. </fontfamily>

--===== _-1288068678==_ma=====--

From owner-modality@LISTSERV.ARIZONA.EDU Wed Apr 14 00:18:22 1999
Date: Wed, 14 Apr 1999 00:14:38 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: possible worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

A few quick thoughts on possible worlds.

In effect, Rachael and Brad bring up the question of how we are to define the notion of "spatiotemporally related". The problem is that if there are counterfactual worlds where I am doing something else now, it seems that my counterpart is in some sense temporally related to me -- he exists at the same point in time, after all. (Some philosophers, and maybe even Lewis, would deny that we can have this sort of "transworld identity" between points in time, just as they deny transworld identity for individuals, but I leave that point aside for now.) Intuitively (as Brad suggests), there is a stonger notion of spatiotemporal relatedness, such that for two entities to be spatiotemporally related they have to be in the same world, so that my counterpart won't count. But Lewis clearly can't just define spatiotemporal relatedness that way, as he wants to define "being in

the same world" in terms of spatiotemporal relatedness, so the definition above would lead to a circle. So the question is whether we can find an independent definition of the stronger notion.

One thing to try might be defining the relation in terms of the existence of a continuous path from A to B. Another thing would be to import causal relations here. A third thing would be to take this sort of spatiotemporal relatedness as a primitive, as e.g. some do on the relational theory of spacetime. On such a theory, two beings or events in the same world might stand in the relation of e.g. A preceding B, or A being a certain distance from B, where these are a primitive sort of relation that only holds, intuitively, "within" a spacetime manifold. But maybe there are better suggestions here. All ideas are welcome.

I think Rachael is write that on Lewis's view, the epistemology of possible worlds becomes more problematic. The analogy with mathematics works better for the ersatz theorist, as we're invoking abstract objects in each case. It's by no means obvious how we can have knowledge of abstract objects without causally interacting with them, but at least there is an intuition here that it's not entirely unreasonable. But in all other cases of concrete objects, it seems that knowledge requires something like causal interaction, or some other intimate epistemic acquaintance.

In response, Lewis argues (pp. 110-12) that the relevant distinction isn't that between the concrete and the abstract, but rather that between the contingent and the necessary. Knowledge of the contingent requires some sort of acquaintance, but knowledge of the necessary doesn't. So I guess the debate comes down to which of these demarcations is most relevant for isolating the sort of knowledge that seems most clearly to require causal (or other) acquaintance. Any thoughts on Lewis's discussion here are welcome.

Re Josh on Lewis on parsimony: I think these are good points. In response, Lewis might say that anyone who is serious about modality will need modal principles or "laws" governing logical space, so he isn't any worse off. Josh might say: it's one thing to have such laws governing abstract objects, another to have them governing concrete objects. Lewis might say: what's the big deal about concreteness? A modal law is a modal law. OK, I need modal laws governing concrete things, whereas you need modal laws governing abstract things, but that doesn't put us in obviously different boats. In any case, we already have modal laws governing concrete things (e.g. the principle of non-contradiction governs concrete things in the actual world).

I think I tend to share Josh's intuition here, but it's a complex dialectic.

Josh's second point is that Lewis needs concrete entities we're not spatiotemporally related to, and that's something new. I think Lewis would agree, but argue that this doesn't involve any new *fundamental* sort of thing. After all, we both have concrete objects, and spatiotemporal relations. Lewis just says something different about when objects stand in those relations. That's a nonfundamental "something new" analogous to me holding that there are swans in Australia where you deny it. And Lewis might argue that the principle of parsimony really on applies to fundamentals, rather than to these complex combinations.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Apr 14 12:00:19 1999
Date: Wed, 14 Apr 1999 11:56:59 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: Re: possible worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

First, I want to give a brief response to Dave on the parsimony issue. I don't think Lewis can say that modal laws are modal laws be they laws for concrete or abstract objects. If possible worlds are concrete then modal laws exist independent of any one possible world. They are in some sense external to the worlds themselves. On the other hand, if possible worlds are abstract objects (or fictions, etc.) then the laws governing them will be laws within the actual world. Furthermore, there will presumably be other laws governing abstract objects of other sorts (ie. numbers), so modal laws will not be a new kind of thing.

Turning to yesterday's discussion, I'm unclear exactly what the difference between and inscrutable truth and an open inconceivability is. Both, I take it, are cases that are negatively conceivable but not positively conceivable. Both are possible. Perhaps open inconceivabilities are things that are to weird to have a concept of, but I don't think we want to use "weirdness" as a formal distinction (though I admit it is an informal distinction). So what did I miss?

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 13 01:20:52 1999
X-Priority: 3
X-MSMail-Priority: Normal
X-MimeOLE: Produced By Microsoft MimeOLE V4.72.3110.3
Date: Tue, 13 Apr 1999 01:18:33 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Brad Thompson <bradt@U.ARIZONA.EDU>
Subject: Re: possible worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

In response to one of Rachael's questions:

>I have two questions- In seminar we were talking about other possible
>worlds which have the same history or are on the same time as our world
>(for example, consider a world where all the history up to this point is
>exactly identical with ours, only this seminar never took place.) It seems
>that this world is spatio-temporally related to ours? If so, should we
>consider it as another possible world or part of this world?

I don't think that we should say that this other world is spatio-temporally related to ours. For one, that other world does not appear to be spatially related to this one. Further, I think it's best to restrict temporal relatedness to within-world events. I remember there was debate about this in seminar. But I don't find any problems with this restriction, nor does it really make sense to me to say that two events that are in different worlds occur at the same time in some absolute sense. Rather, when we

describe two worlds that have the same history up until a particular time (within each world), events across worlds that we want to describe as "happening at the same time" ought to be described rather as occupying symmetrical locations in their respective world-histories.

Brad

From owner-modality@LISTSERV.ARIZONA.EDU Fri Apr 9 19:04:01 1999
Date: Fri, 9 Apr 1999 19:02:47 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Next week, etc.
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

For next week's meeting we are focusing on modal rationalism. This will be the culmination of the current phase of the course. In the first half of the meeting we'll go over the "master argument" for modal rationalism (and against modal dualism) as presented in section 3.5 of "Materialism and the Metaphysics of Modality" and section 2.9 of "Mind and Modality". In the second half we will tie up some loose ends re inscrutability, open inconceivabilities, etc (corresponding roughly to 2.6 of "Mind and Modality"), and look at the general shape of modal rationalism.

In the following week, we'll return to some issues concerning the two-dimensional semantic framework. First, we'll finally get to the material on "the tyranny of the subjunctive", concentrating on links to issues concerning subjunctive and indicative conditionals. Second, we'll look at the application of the framework to issues concerning the contents of thought, and in particular issues concerning internalism, externalism, and narrow content (the central reading will be my paper "The Components of Content"). Third (in what I think will be the final week), we'll look at some objections to the 2-D framework as an account of the content of thought, e.g. those raised by Block and Stalnaker (separately and together), and will look at some of the intricacies, limitations, and open questions about the framework.

In the meantime, I look forward to seeing everyone's discussion of the material on possible worlds that we covered this week.

Finally, there is a party tomorrow (Saturday) night at Stuart Hameroff's house, which is where I'm staying. It will go from 7pm onward, and beer, wine, and food will be provided. The majority of the people there will be doctors, nurses, and such from the hospital (Stuart is an anesthesiologist who dabbles in consciousness research), but there will be a few philosophers and such there. The house is at 2144 E. Camino El Ganado, in the foothills. To get there: go north on Campbell from River for a couple of miles until just after Campbell "divides", turn left on Camino La Zorella (the second left after the divide), left again on Camino Escalante, and right on El Ganado. The house is on the corner; easiest to use the second driveway on the left. Sorry about the short notice -- hope you can make it!

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 13 11:53:34 1999
Date: Tue, 13 Apr 1999 11:40:30 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik A Herman <erikh@U.ARIZONA.EDU>
Subject: Re: possible worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

I'm having some worries about a priori conceivability. How do we remove the a posteriori influences from our a priori concepts? I'm not sure I would have come upon operators like addition, or conjunction, or negation,

or number, if I didn't see the world first. How can we tell that the effects of experience haven't influenced the data here?

Also, if we were to have used modal rationalism to help hypothesize physical truths in science, would we have ever come up with quantum mechanics? They defy what seem logically possible (annihilation and creation of particles & immediate effects across space).

Dave, I don't understand the part where you say that "we need the modality of ideally conceivable worlds = possible worlds...even if some of these worlds aren't metaphysically possible." And then in the next paragraph, "there is no bar to the existence of logically possible worlds." (Mind and Modality 2.9)

-Erik H.

From erikl@U.Arizona.EDU Fri Apr 16 15:59:35 1999
Date: Fri, 16 Apr 1999 15:58:34 -0700 (MST)
From: Erik J Larson <erikl@U.Arizona.EDU>
To: David Chalmers <chalmers@LING.UCSC.EDU>
cc: MODALITY@LISTSERV.ARIZONA.EDU
Subject: Re: Type-A possibilities
Status: RO

I've been thinking lately about the possibility of a type-A defense of materialism. Here's why. Given that a type-b defense a) requires some conceptual analysis (eg apriori reasoning) to connect physical and phenomenal states that are prima facie under different descriptions, and b) such conceptual analysis, if one grants the cogency of the 2D framework, will not deliver strong necessities (necessary aposteriori statements with necessary primary intensions) but rather more Kripkean-like examples (water = H₂O) that have contingent primary intensions, then c) the type-b position will end up postulating a space of metaphysically possible worlds to deny the possibility of zombies in order to break the conceivability to possibility link, whereby it (materialism) will run aground by biting the bullet of modal dualism--having as it were no good motivation on its behalf other than its consistency with the antecedent claim that materialism must be true. So, that is a lot to say, and a lot to argue about, and I want to just grant it all--whatever more may be said of it or against it--and consider the type-a position here.

In the type-a, I take it that there is an apriori necessity (and hence a necessary primary intension), because there are not distinct sets of terms that fall under different descriptions, and so the explanatory gap is illusory (and temporary). Consider two claims 1) Phenomenal stuff doesn't exist and only physical stuff does (in the sense that an eliminativist means), and 2) Phenomenal stuff does exist. This is something like Goldbach's conjecture in at least this respect--if it turns out that 1) is true then it would seem that it is necessarily true, since the claim would then come down to evaluating $P = P$ by substitution with $P(\text{physical}) = Q(\text{phenomenal})$, which is necessarily true (physical states are equal to themselves necessarily), in which case 2) would become necessarily false. Then we would be *misdescribing* a possible world where 2) held, since we could not be conceiving of a world where something that was necessarily false was in fact the case.

Now, it seems that the legitimacy of the type-a position comes down to whether we can cash out the truth, and hence necessity, of 1) apriori. This amounts to getting an apriori "bridge" between apparently different states such that we can see phenomenal states as physical states (I think this is close to something a type-a would hold). Churchland, et al (ie, Churchland) motivates this possibility by appealing to the progress of

science changing our concepts of things (roughly, the way we conceptualize the world and our experiences), which amounts to a "wait and see" reply to claims of the antecedent absurdity of supposing the primary intensions of physical and phenomenal states may come together some day. But why not? What IF the formation of our primary intensions of physical and phenomenal states are themselves contingencies that are subject to change and alteration; alteration that we as of yet cannot envision, but that remains nonetheless an open question to be filled in, say, by a completed neuroscience? To be somewhat more speculative, what if a completed cognitive science discovered ways in which we systematically misuse concepts to create illusions of separation between actually, ideally, unified phenomena? Might THIS science be in a position to provide a road map for creating the apriori bridge between physical and phenomenal intensions? And if this is at least possible, might materialism be true after all?

Erik

Some advice for drivers on the University of Arizona campus:

"When passenger of foot heave in sight, tootle the horn. Trumpet him melodiously at first, but if he still obstacles your passage then tootle him with vigor."

--From a brochure of a car rental firm in Tokyo

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Mon Apr 19 23:43:00 1999
Date: Mon, 19 Apr 1999 23:42:48 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Type-A possibilities
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: RO

Erik L. canvasses the possibility of a sort of type-A materialism. In terms of the later taxonomy in which type A split into two parts (type A and type C), Erik's suggestion is a sort of type-C materialism, with an a priori connection that is not obvious to us yet, but which may become obvious on some sort of conceptual development.

>What IF the formation of our primary intensions of physical and phenomenal
>states are themselves contingencies that are subject to change and
>alteration; alteration that we as of yet cannot envision, but that remains
>nonetheless an open question to be filled in, say, by a completed
>neuroscience? To be somewhat more speculative, what if a completed
>cognitive science discovered ways in which we systematically misuse
>concepts to create illusions of separation between actually, ideally,
>unified phenomena? Might THIS science be in a position to provide a road
>map for creating the apriori bridge between physical and phenomenal
>intensions? And if this is at least possible, might materialism be true
>after all?

Interesting. The suggestion is that our primary intensions themselves may change over time. In response the anti-materialist might make various points. E.g. (i) it's hard to find cases in science where

primary intensions change a great deal -- do you have some analogous cases in mind? Or (ii) in cases where primary intensions do change, one can make the case that the change is largely terminological, and that there's an alternative history where the word stayed with the old primary intension, and perhaps a new word was developed for the new intension.

But perhaps most important is (iii), if the old term really did have a primary intension, then *that* primary intension determines a perfectly good property (function from worlds to individuals), and presumably the anti-materialist argument will still work for *that* property. Perhaps the term later comes to have a new intension that the argument doesn't work for, but that doesn't stop the argument working for the old intension.

One can imagine something like this happening e.g. if the PI of "consciousness" gradually moved from a purely phenomenal notion to a functional notion (as might well happen as a matter of sociological fact). The later functional PI won't support an anti-materialist argument, but nevertheless the earlier phenomenal PI will. As long as (i) the early term has a PI not entailed by the physical facts, and (ii) the statement "consciousness exists" or some such, with the early term, expresses a truth, then the argument will still go through.

When I visited Alvin's seminar on consciousness last spring, one of the students (Kurt, I think) read a very interesting paper that defended more or less the line you're suggesting. He argued that the PI of some concepts (e.g. "life") has changed, and that a changed PI for "consciousness" could save materialism. In response I argued that (i) it's not so clear that the PI of "life" (as opposed to some beliefs about it) has changed, and (ii) a simple change in PI isn't enough to do the job for the materialist, for the reasons outlined above.

So it seems to be best for the type-C materialist not to say that the PI will change, but rather that the PI of our concept will turn out to be other than we think it is. E.g., perhaps further a priori reasoning will reveal that there is a deep a priori functional analysis of our concept of consciousness, and we just haven't figured that out yet. If so, then the PI of our concept has been functional all along. We just haven't done all the reasoning required to bring that out.

Maybe the line you're suggesting could in fact have that consequence. It's tricky because the "cognitive science" part is presumably empirical, whereas what's required for a PI is a priori. Still, the suggested claim that we "misuse" concepts suggest a failure of rationality on our part, one that could be corrected by more ideal reasoning. One would hope that we could detect that failure within the a priori domain, though. If all that's going on is the Loar-style point that we have distinct concepts, apriori-unconnectable, that nevertheless refer to the same physical stuff, then we have type-B materialism. I presume that's not what you want. But to get the difference, it's crucial that the concepts be connectable in the a priori domain.

Of course, to get that connection in the a priori domain, we will need to figure out that deep down there is a structural or functional analysis of the concept of phenomenal consciousness. I've given reasons to be skeptical of that claim; on the face of it, such an analysis is just changing the subject. But the type-C materialist could always argue that the apparent failure is because of a deep

cognitive blindness on our part (shades of McGinn, except here perhaps the blindness is correctable). That would be a tricky case to make, but there's interesting ground to be plowed over here.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 20 12:54:49 1999
x-sender: agillies@pop.u.arizona.edu
Date: Tue, 20 Apr 1999 13:08:51 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: Re: truth conditions and conceivable worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Here's a bit of an exchange from last week about 2D semantics. I think I have a conjured sentence that fits the bill. I'll put it below.

>>This points to another confusion I've been entertaining on 2D semantics.
>>It's probably a little off topic, so we'll get to it in a week or two.
>>But the main idea is to what extent the 2D semantics for terms meshes
>>with the 2D semantics for sentences. In particular, in figuring the
>>truth₁ of a sentence S, it looks like every term in S must be mapped to
>>its referent₁. Likewise for truth₂. Are there sentences S for which
>>the natural truth₁ conditions might depend on some term in S being
>>mapped to its referent₂?

>
>Hmm, interesting. Again, a statement is true₁ at the actual world
>iff it is true₂ there. And a term's (actual) referent₁ will always
>be its (actual) referent₂ -- both are just the term's actual
>referent! It's only in non-actual possible worlds that these things
>come apart. But I suppose your suggestion might come to the
>possibility that to evaluate the truth₁ of S at some non-actual
>possible world, we need to consider the referent₂ of some term in S
>there, which would then depend on a posteriori facts about the actual
>world.

>
>The short answer, I think, is that this can't happen. Truth₁ is
>defined so that it depends only on facts about the world in question
>(plus a priori analysis). Some statements, e.g. "The actual president
>is Bill Clinton" might make reference to the actual world, but they
>can still be evaluated in a single-world way. E.g. to evaluate the
>truth₁ of the statement above at a world where George Bush is
>president, we consider that world as actual, and determine that if
>that world is actual, the statement is false. That's to say, one
>never needs to import the *actual* referent of "the actual president"
>in evaluating the truth₁ of this statement at world W. One only ever
>imports what the referent would be *if* W were actual.

>
So here's the sentence:

(\$) The water in The River Thames is H₂O.

What I want to know is whether this is necessary or not where what we are interested in is how "water" is evaluated in worlds considered as actual, but where "The River Thames" is given the semantic value when considered counterfactually. That is, in computing the semantic value of (\$) we want the referent given by the PI of "water" but the SI of "The River Thames". If I'm right, this isn't possible on the 2D account. So there is a space of sentences for which it isn't defined---namely, any sentence where we "cross" intensions of the terms.

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 20 15:28:32 1999
Date: Tue, 20 Apr 1999 15:22:25 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: truth conditions and conceivable worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

>So here's the sentence:

>

>(\$) The water in The River Thames is H₂O.

>

>What I want to know is whether this is necessary or not where what we are
>interested in is how "water" is evaluated in worlds considered as actual,
>but where "The River Thames" is given the semantic value when considered
>counterfactually. That is, in computing the semantic value of (\$) we
>want the referent given by the PI of "water" but the SI of "The River
>Thames". If I'm right, this isn't possible on the 2D account. So there
>is a space of sentences for which it isn't defined---namely, any sentence
>where we "cross" intensions of the terms.

Hmm, I'm not quite sure what the counterexample sentence is supposed
to be. If the sentence is just (\$), then it isn't a counterexample,
as this has a perfectly straightforward PI and SI. Rather, I take it
your counterexample is some modification or "interpretation" of (\$), on
which we "give the terms semantic values" according to their PI and SI
respectively in different cases. It's not clear to me in what sense
the result is a "sentence".

Maybe given a sentence with a well-defined PI and SI, one can define
some abstract object which gets evaluated in possible worlds according
to some combination of the PIs and the SIs in the original sentence.
But this doesn't seem to correspond naturally to anything we find in
language. It doesn't seem to be meaningful to talk about the PI or SI
of such an abstract object; but that alone isn't obviously a problem
for two-dimensional semantics. 2-D semantics is meant to apply to
sentences in natural languages, not to arbitrary abstract objects.

It seems to me that on any natural reading of (\$), it gets evaluated
in worlds considered as actual according to the PI of both "water" and
"River Thames" -- one can't just stipulate that it will be evaluated
according to the SI of "River Thames". At least, if one did, one
would have a very different sentence where the word "River Thames"
means something different to what it means in ordinary English (a new
word whose PI happens to be the same as the SI of our word "River
Thames"). Even if one did this, one could now naturally evaluate the
PI and the SI of the new sentence in a pretty straightforward way; it
would have a PI that differs from that of the original sentence, but
that's only to be expected.

I may be missing the way you intend the example to be understood, in
which case I'd be interested to hear it articulated.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Wed Apr 21 00:15:08 1999
Date: Wed, 21 Apr 1999 00:14:19 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>

Subject: Next week
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Dear all,

For next week we are focusing on issues re the contents of thought. In particular, we will be focusing on issues about internalism, externalism, and "narrow content", and the application of the 2-D framework to these issues.

The central reading is my paper "The Components of Content". You've read this already earlier in the term for the general issues re 2-D semantics; it might be worth reading it again for the more specific issues concerning the narrow content, the explanation of behavior, belief ascription, and so on.

There are also a number of background readings that Jodie will photocopy and put in the folder, hopefully tomorrow. These are:

- Putnam, "The Meaning of 'Meaning'"
- Burge, Excerpt from "Individualism and the Mental"
- White, "Partial Character and the Language of Thought"
- Stalnaker, "Narrow Content"
- Schiffer, "The Mode-of-Presentation Problem".

The Putnam and the Burge are the classic sources re externalism. Many or most of you will have read them before. If you haven't, I certainly recommend reading them; if you have, you might look over them again. The White and Stalnaker are two papers on whether we can make sense of a notion of "narrow content" (a sort of content that is "in the head") in response to the externalist arguments. White ends up arguing for a picture that has some similarities to the 2-D framework, while Stalnaker argues against narrow content and on the way argues against the use of the 2-D framework. Both also have some good general discussion. Finally, the Schiffer has good coverage on some of the issues re the semantics of belief ascription that come up in my content paper.

The Stalnaker will be more central next week (when we will discuss objections to and problems for the 2-D framework), but you might find it useful to look over this week for its general way of framing the issues. I'm not sure whether we will discuss the White in detail (it was a special request from Brad), but again it is useful and interesting background material. The same goes for Schiffer. Hopefully these papers will help make some of the background to "The Components of Content" fall into place.

Finally, the folder will also have a copy of John Burgess's article "Quinus ..." (I've forgotten the full Latin title, but it translates as "Quine Freed of Every Flaw"). That's the article I was discussing today re the debate over quantified modal logic, with some reference to subjunctive and indicative conditionals. This isn't an official reading, but it's a very interesting and enlightening discussion of the issues re quantified modal logic (it should also work well as background for those who don't know much about the debate), and I recommend it to everyone.

A reminder that I am going to England tomorrow, and will be arriving back in town next Tuesday at 4pm, so next week's meeting is scheduled for 5:20pm Tuesday. I'll get in touch by e-mail or phone if it looks like I'll be late.

Also, as I said in class today, anyone who wants to give me a draft of their term paper by next Friday (April 30) is welcome to do so, and I'll give them comments back.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 22 16:32:44 1999
Date: Thu, 22 Apr 1999 16:30:55 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Re: truth conditions and conceivable worlds
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Regarding Dave's response to Thony's "River Thames" sentence, it strikes me that the 2D framework is immune to the possibility that a space of sentence may have uncomputable truth values (due to mixing intensions), but for a reason that may itself raise a red flag. It's just been *stipulated* that statements get evaluated according to primary or secondary intensions, uniformly for all terms. That keeps Thony's example from gaining any ground, I suppose, but given that certain statements are most naturally interpreted by assigning different intensions (take Thony's example, for one), the stipulation cries out for justification, at least as a general theory of semantics. I suppose the justification is practical, or *methodological*, in that we can apparently put the 2D semantics to good use (eg, in evaluating Kripkean cases). But given that we can't make sense of "The water in the River Thames is H₂O" by taking the natural intensions of the composite terms, it seems somehow ad hoc. Part of my frustration here stems from my own attempt to formulate statements whose natural interpretation mixes intensions, until I realized (from reading Dave's message) that it's just been *stipulated* that you can't do that. (Trivially, I managed to come up with statements that always end up false when evaluated according to the PI, even when the composite clauses of the sentences get different truth values when evaluated in different worlds considered as actual. Consider "Water is actually H₂O, but not here." This is equivalent to the conjunction "Water is actually H₂O and Water is not H₂O here.", where the first conjunct is false and the second true in Twin Earth, and vice versa for Earth. (Since conjunction are false when at least one conjunct is false, the statement is always false no matter in what world considered as actual it is evaluated).) This I say is trivial, because who cares if we can construct sentences contradictory or otherwise according to an intension, as long as our semantics always *gives* us a truth value one way or the other. Thony's example was meant to challenge the latter, but it evidently can't be done, in the same sense that you won't likely find any round squares laying around either. Questions: How seriously do we take the 2D framework as a theory of semantics? Should we be bothered by its inflexibility with respect to Thony's example? To be really speculative, might *adding* the sort of flexibility of evaluation that can produce compositional semantics using different intensions from the terms on up possibly *undermine* the motivation for the anti-materialist argument, but give us new methodological justification for its acceptance nonetheless (because it is a more flexible semantics)?

On Tue, 20 Apr 1999, David Chalmers wrote:

> >So here's the sentence:
> >

> >(\$) The water in The River Thames is H₂O.
> >
> >What I want to know is whether this is necessary or not where what we are
> >interested in is how "water" is evaluated in worlds considered as actual,
> >but where "The River Thames" is given the semantic value when considered
> >counterfactually. That is, in computing the semantic value of (\$) we
> >want the referent given by the PI of "water" but the SI of "The River
> >Thames". If I'm right, this isn't possible on the 2D account. So there
> >is a space of sentences for which it isn't defined---namely, any sentence
> >where we "cross" intensions of the terms.
>
> Hmm, I'm not quite sure what the counterexample sentence is supposed
> to be. If the sentence is just (\$), then it isn't a counterexample,
> as this has a perfectly straightforward PI and SI. Rather, I take it
> your counterexample is some modification or "interpretation" of (\$), on
> which we "give the terms semantic values" according to their PI and SI
> respectively in different cases. It's not clear to me in what sense
> the result is a "sentence".
>
> Maybe given a sentence with a well-defined PI and SI, one can define
> some abstract object which gets evaluated in possible worlds according
> to some combination of the PIs and the SIs in the original sentence.
> But this doesn't seem to correspond naturally to anything we find in
> language. It doesn't seem to be meaningful to talk about the PI or SI
> of such an abstract object; but that alone isn't obviously a problem
> for two-dimensional semantics. 2-D semantics is meant to apply to
> sentences in natural languages, not to arbitrary abstract objects.
>
> It seems to me that on any natural reading of (\$), it gets evaluated
> in worlds considered as actual according to the PI of both "water" and
> "River Thames" -- one can't just stipulate that it will be evaluated
> according to the SI of "River Thames". At least, if one did, one
> would have a very different sentence where the word "River Thames"
> means something different to what it means in ordinary English (a new
> word whose PI happens to be the same as the SI of our word "River
> Thames"). Even if one did this, one could now naturally evaluate the
> PI and the SI of the new sentence in a pretty straightforward way; it
> would have a PI that differs from that of the original sentence, but
> that's only to be expected.
>
> I may be missing the way you intend the example to be understood, in
> which case I'd be interested to hear it articulated.
>
> --Dave.
>

Some advice for drivers on the University of Arizona campus:

"When passenger of foot heave in sight, tootle the horn. Trumpet
him melodiously at first, but if he still obstacles your passage
then tootle him with vigor."

--From a brochure of a car rental firm in Tokyo

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 27 11:49:56 1999
Date: Tue, 27 Apr 1999 11:48:10 -0700
Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik A Herman <erikh@U.ARIZONA.EDU>

Subject: Intuitions on Components of Content

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: R

I would like to resolve my intuitions with the 2-Framework and I'm not quite sure how to do it. I've listed my intuitions on each of the 6 cases:

1) The twins relation to their environment is identical.

2) There is a difference between saying "Hesperus is Hesperus", and "Hesperus is Phosphorus". The first is just a figurative way of saying "a thing is itself", which HAS meaning, just not a very important one because it's so trivial and obvious. The second, on the other hand, is more like saying "two things are the same when you think about it". Content doesn't just denote something's physical embodiment but it's entire meaning. "the former is trivial and the latter is not"

3) Kripke's puzzle happens in real life. And the language issue doesn't do anything but explicate that Pierre doesn't see that THAT'S the "London" (however you wanna say it) that he's been told is pretty. Consider this real life scenario: I was told in phone conversations and e-mails "Mike is a cool guy." Then one night I was at a party and I met a guy, his name was Mike. Later, my friend asked if I had met Mike at the party, and being the flaky guy I am, said, "Well yes, I met a guy named Mike, why?" and she responded "THAT'S MIKE!" (meaning, that's THE Mike she had been speaking of in our correspondence). At which time I "linked/combined" these two, otherwise separated references. Now there doesn't seem to be anything confusing about this, and yet it is similar, in the relevant way, to the Londres case. If I had thought Mike was a jerk when I met him, this contradiction would be resolved in the "linking/combining" stage--analyzing my reasons for thinking he's a jerk with her reasons for thinking he's cool and weighing everything in the mix. I might even keep BOTH contents because they differ in their relation to Mike. To make the analogy hold, we could say that at the party he introduced himself as "Michel" or something stupid like that but this doesn't do any work.

4) My favorite video game is a car race (I forget what it's called) that is at the Flying J truckstop halfway to Phoenix (that's relevant). When I play it and I believe that "I'm" going to crash into a guardrail, I take evasive action. After I'm done playing, I get into the real car and I get yelled at by my girlfriend because I'm apparently driving like I'm still in the video game. The distinction is made in the differing degrees of indistructability. There is no difference IN KIND between the case where I am indestructible and the case where I might actually get hurt, just a severe difference in degree of personal meaning they have. The same concept holds for the difference in degrees of recklessness in driving a car -vs- a motorcycle.

5) "Clark Kent" and "Superman" have two totally different meanings, the only similarity is that they occupy the same body.

6) This one has me stumped.

-erik h.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 27 13:00:54 1999

Date: Tue, 27 Apr 1999 12:56:54 -0700

Sender: "Philosophy 596B: Mind and Modality"

<MODALITY@LISTSERV.ARIZONA.EDU>

From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Components of Content
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

(1) First, a point about Thony's and Erik L's comments on why mixed intension sentences are naughty. If we take *concepts* as primary, then the strictures against mixed intension sentences would seem to be ad hoc. If, on the other hand, we take *thoughts* as primary, then it might be reasonable to hold that mixed intension thoughts are a no-no, b/c they're not really thoughts. Problem is, most semantics are compositional: take concepts as primary, and build up thoughts out of them. Reply: sure semantics is compositional, but even so, <Green ideas sleep furiously> doesn't make any sense (or, doesn't give one a coherent thought), as Thony would no doubt be the first to agree. In the same way, <The water [1-intension] in the Thames [2-intension] is H2O [1-intension]> or what-have-you, is also nonsensical. So, we're going to have to give some account of why <Green ideas sleep furiously> is not a legitimate thought, why can't we just plug in this account to rule out mixed intension sentences?

(2) I'm losing sleep over the issue of whether content is effable. (There are two ways in which one might cash effability out: thinkability, communicability/sayability. I'm going to stick with the latter.) Fodor says that his notion of narrow content is ineffable. I'm worried that *both* of Chalmers's notions of content are ineffable. There are different ways to put my worry and I'm not sure which is best, but here's a stab.

Sometimes the account of notional/relations content sounds as if there are actually two distinct concepts per 'folk' concept or word. Or, to put it at the level of thoughts, if thoughts are individuated by their truth conditions, and if 'a' thought interpreted via 1-intension has one truth-condition and 'the same' thought interpreted via 2-intension has different truth-conditions, then there are different thoughts. Thus, it would be possible to have a language in which 1- and 2-intensions are explicitly distinguished at a syntactic level. In the same way that some languages build temporal discriminations into verb endings, our hypothetical language builds intensional discriminations into its words. *Indeed, I take it that this is just what we've done in this seminar when we write "water 1-intension".* Further, I take it that Chalmers sees cognitive psychology as moving towards a more purely notional/1-intensional language. (p. 12)

BUT on the other hand, there is evidence that one can not, even in principle, prise the notional apart from the relational in this way. The problem is that if we split folk concepts into two concepts, the notional and the relational, then *THERE IS NO END TO THIS PROCESS, FOR EVERY CONCEPT HAS A PRIMARY INTENSION AND A SECONDARY INTENSION.* In other words, it looks like you're going to get something like a 'what-Achilles-said-to-the Tortoise' kind of problem. Tortoise: "What is the 1-intension of X?" Achilles: "X, Y and Z". T: "Do you mean X, Y and Z in the 1-intension sense, or the 2-intensions sense?" A: "The 1-intension sense". T: "Ah, and that would be?" T: "F, G and H". T: "Ah, do you mean F, G and H in the 1-intension sense, or the 2-intension sense?". and so on.

Chalmers says that 'we should not mistake the linguistic expression [of 1-intensions] for the real thing' (p. 13) Fair enough, but if there is a real thing here, should we be able to *unambiguously* refer to it. And it looks like we can't, for every attempt that one makes to refer to notional or relational content *via words* (and what else can one use?) can be interpreted in one of two ways, either 1-intensionally or 2-intensionally.

(Actually, I think that there's more than one problem contained in the above, but it's lunchtime and I'm not going to rewrite it.)

(3) I'm worried about the implications Kripke's rule-following has for the accessibility of notional content. Chalmers says that notional content is always (in principle, at least) accessible. It is internal to the thinker. But consider Kripke's plus/quus concerns. It's not exactly clear what these are, but here's a stab: Do I know what I mean by 'plus' merely by introspecting? No. In fact, I only mean something by plus by 'how I go on', that is, by the rule I follow *over time*. So how can I know what I mean by 'plus' *now* if what I do in fact mean is in part fixed by the rule that I follow? In a nutshell, the worry is that (certain) notional contents are fixed by rules that I follow over time, and these may be cognitively inaccessible to me *at a time*.

t.

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 29 00:33:43 1999
Date: Thu, 29 Apr 1999 00:05:00 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Brad Thompson <bradt@U.ARIZONA.EDU>
Subject: Re: Components of Content
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Here are some comments on just some of Tim's message:

>Chalmers says that 'we should not mistake the linguistic expression [of
>1-intensions] for the real thing' (p. 13) Fair enough, but if there is a
>real thing here, should we be able to *unambiguously* refer to it.

We do unambiguously *refer* to primary intensions, simply by saying "the primary intension of x". What seems less clear is whether we can *specify* the primary intension, that is, give a conceptual analyses.

And it
>looks like we can't, for every attempt that one makes to refer to notional
>or relational content *via words* (and what else can one use?) can be
>interpreted in one of two ways, either 1-intensionally or 2-intensionally.

I can't think of two intensions for "the primary intension of 'water'". There can be cases where the primary and the secondary intension coincide, and maybe *complete* conceptual analyses of primary intensions will consist in only these unambiguous statements. If not, can't we just stipulate the intended interpretation like above, by prefacing the statement with "the primary intension of..." or "the secondary intension of..."?

>(3) I'm worried about the implications Kripke's rule-following has for the
>accessability of notional content. Chalmers says that notional content is

>always (in principle, at least) accessible. It is internal to the thinker.
>But consider Kripke's plus/quus concerns. It's not exactly clear what
>these are, but here's a stab: Do I know what I mean by 'plus' merely by
>introspecting? No. In fact, I only mean something by plus by 'how I go
>on', that is, by the rule I follow *over time*. So how can I know what I
>mean by 'plus' *now* if what I do in fact mean is in part fixed by the
>rule that I follow? In a nutshell, the worry is that (certain) notional
>contents are fixed by rules that I follow over time, and these may be
>cognitively inaccessible to me *at a time*.

I do think that you know merely by introspecting what you mean by "plus". I'm not prepared to give a full-scale argument against Kripke though. Would Kripke really say that the meaning of "plus" is fixed by the rule that you follow? This leaves meaning forever indeterminate, since at the time of your death we will still be left with an infinitude of possible rules that you *were* following. It has been awhile since I read it, but I thought that Kripke's resolution of the problem was to go social with meaning. If so, then this can easily be accommodated within the 2d framework. The primary intension of "plus" is [the rule that I am following as specified by my linguistic community] or something like that. This is no different in kind, and no more problematic, than the primary intension of water.

Brad

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 29 15:28:26 1999
Date: Thu, 29 Apr 1999 15:14:13 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Timothy J Bayne <bayne@U.ARIZONA.EDU>
Subject: Re: Components of Content (fwd)
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Here's a reply to Brad's objection to my Kripkean worry, I'm still thinking Brad's reply to my first worry. 1st paragraph is my original, then Brad's comment, then my reply.

>
> >(3) I'm worried about the implications Kripke's rule-following has for the
> >accessability of notional content. Chalmers says that notional content is
> >always (in principle, at least) accessible. It is internal to the thinker.
> >But consider Kripke's plus/quus concerns. It's not exactly clear what
> >these are, but here's a stab: Do I know what I mean by 'plus' merely by
> >introspecting? No. In fact, I only mean something by plus by 'how I go
> >on', that is, by the rule I follow *over time*. So how can I know what I
> >mean by 'plus' *now* if what I do in fact mean is in part fixed by the
> >rule that I follow? In a nutshell, the worry is that (certain) notional
> >contents are fixed by rules that I follow over time, and these may be
> >cognitively inaccessible to me *at a time*.
>
> I do think that you know merely by introspecting what you mean by
> "plus". I'm not prepared to give a full-scale argument against Kripke
> though. Would Kripke really say that the meaning of "plus" is fixed
> by the rule that you follow? This leaves meaning forever
> indeterminate, since at the time of your death we will still be left
> with an infinitude of possible rules that you *were* following. It
> has been awhile since I read it, but I thought that Kripke's
> resolution of the problem was to go social with meaning. If so, then
> this can easily be accommodated within the 2d framework. The primary
> intension of "plus" is [the rule that I am following as specified by
> my linguistic community] or something like that. This is no different

> in kind, and no more problematic, than the primary intension of water.

I think Brad is probably right in thinking that Kripke wants to locate the rules in social practices, but this cannot be accommodated with the framework as presented in "The Components of Content", b/c *I* don't have direct access to the rules of my community. The problem now isn't one of my having access to a diachronic fact at this moment, but *my* having access to a fact that is fixed by social practices. How do I know what rule the rest of the folks in my community are now following? Wouldn't that be like saying that I have direct introspective access to the fact that my community thinks that it is rude to burp at the table? If what I now mean is fixed by socially constituted facts, how do I know what I mean?

t.

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From owner-modality@LISTSERV.ARIZONA.EDU Thu Apr 29 15:43:32 1999
Date: Thu, 29 Apr 1999 15:23:14 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: David Chalmers <chalmers@LING.UCSC.EDU>
Subject: Re: Components of Content
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Re effability of notional content, and re Kripke on rule-following.

Tim writes:

>(2) I'm losing sleep over the issue of whether content is effable. (There
>are two ways in which one might cash effability out: thinkability,
>communicability/sayability. I'm going to stick with the latter.) Fodor
>says that his notion of narrow content is ineffable. I'm worried that
>*both* of Chalmers's notions of content are ineffable. There are different
>ways to put my worry and I'm not sure which is best, but here's a stab.
>
>Sometimes the account of notional/relations content sounds as if there are
>actually two distinct concepts per 'folk' concept or word. Or, to put it
>at the level of thoughts, if thoughts are individuated by their truth
>conditions, and if 'a' thought interpreted via 1-intension has one
>truth-condition and 'the same' thought interpreted via 2-intention has
>different truth-conditions, then there are different thoughts. Thus, it
>would be possible to have a language in which 1- and 2-intensions are
>explicitly distinguished at a syntactic level. In the same way that some
>languages build temporal discriminations into verb endings, our
>hypothetical language builds intensional discriminations into its words.
>*Indeed, I take it that this is just what we've done in this seminar when
>we write "water 1-intension".* Further, I take it that Chalmers sees
>cognitive psychology as moving towards a more purely
>notional/1-intensional language. (p. 12)
>
>

>BUT on the other hand, there is evidence that one can not, even in
>principle, prise the notional apart from the relational in this way. The
>problem is that if we split folk concepts into two concepts, the notional
>and the relational, then *THERE IS NO END TO THIS PROCESS, FOR EVERY
>CONCEPT HAS A PRIMARY INTENSION AND A SECONDARY INTENSION.* In other
>words, it looks like you're going to get something like a
>'what-Achilles-said-to-the Tortoise' kind of problem. Tortoise: "What is
>the 1-intension of X?" Achilles: "X, Y and Z". T: "Do you mean X, Y and Z
>in the 1-intension sense, or the 2-intensions sense?" A: "The 1-intension
>sense". T: "Ah, and that would be?" T: "F, G and H". T: "Ah, do you mean F
>, G and H in the 1-intension sense, or the 2-intension sense?". and so on.

OK, here's what I think. First, I do want to say that it is *one*
concept that has both a primary intension and a secondary intension.
E.g. a being might have just the concept of "water" and not the
concept of "H2O". So they'll have just one concept in the vicinity,
with both a primary intension and secondary intension. The primary
intension will pick out watery stuff in all worlds, and the secondary
intension will pick out H2O in all worlds, but that is not to say the
being has separate concepts of "watery stuff" and of "H2O". There is
just the one "water" concept, with a complex structure of semantic
evaluation.

Of course it will often happen that when a being has a concept with a
primary intension and a secondary intension, that being will have two
concepts in the vicinity which come close to mirroring each of these.
For example, as a matter of fact, in addition to my concept of "water"
I have a concept of "watery stuff" and "H2O". It isn't required that
I have these distinct concepts to have the concept of "water", though.

In effect, what is going on in those cases is that one has an initial
concept C1 (e.g. "water") with $PI(C1) = P$ and $SI(C1) = S$. Then one
finds a further concept C2 (e.g. "watery stuff") such that $PI(C2) = P$
and $SI(C2) = P$. That is, C2 is a concept such that *both* its PI and
its SI mirror the PI of C1 -- so we can see C2 as in some sense
"isolating" the PI of C1 more purely.

And then one finds a still further concept C3 (e.g. "H2O") such that
 $PI(C3) = S$ and $SI(C3) = S$. That is, C3 is a concept such that *both*
its PI and its SI mirror the SI of C1 -- so we can see C2 as in some
sense "isolating" the SI of C1 more purely.

All this is nice to "articulate" the 2-D structure of the initial
concept P1. But it certainly isn't required just to have the initial
concept P1. To do that, one needs just one concept. It's the
articulation that requires further concepts.

Note that when we "articulate" a concept in this way, your
Achilles/Tortoise problem doesn't quite come up, as the further
concepts we've introduced have more or less the same PI and SI. Or at
least, they are closer to doing so than the original concept.

Still, it's not obvious to me to what extent we can really expect the
PI and SI of our concepts to be "articulated". As far as I can tell,
that more or less requires a reductive definition of either of these
intensions in terms that don't involve the original concept. E.g., we
want to define the PI and SI of "water" without appealing to water.
Now that is sometimes possible and sometimes isn't, as philosophers
well know from experience. But it's not clear to what extent such
definability is really required to say that content is "effable".
Take the concept of "knowledge" -- maybe not definable, but do we want
to say it is ineffable? I don't think the PI and SI of our concepts

in general will be any worse than this.

Why do people sometimes say that the wide content of a concept such as "water" is effable? Perhaps because one can say something like: the SI of "water" picks out the water in all worlds. But that obviously isn't all that helpful -- one is just implicitly appealing to the SI to characterize the SI. And one could do something similar to characterize the PI.

Alternatively, maybe they think the content here is effable because one can say the SI of "water" picks out H₂O in all worlds. Here we have something like a reductive definition. But (a) it's not clear this is possible for all concepts (it's not obvious that it really even works for "water"), and (b) something similar can work for primary intensions, via e.g. "watery stuff". Of course what goes on is, in effect, that we find another concept ("H₂O") which has the same SI as "water", and whose PI is a lot closer to that SI too. So similarly, with the PI, we find another concept ("watery stuff") which has the same PI as "water", and whose SI is a lot closer to that PI too. That seems a legitimate thing to do.

Whether this sort of reductive definability is possible for all concepts, who knows? My own view is that it probably isn't possible for some central concepts, such as e.g. concepts of experience, cause, space, and time (cf. discussions of global Ramsification in Jenann's seminar). For other concepts (e.g. knowledge), approximations at reductive definitions may be possible, but not perfect definitions. And for others (e.g. bachelor), there may be near-perfect definitions.

It seems to me that this was the position that we knew we were in before thinking about these things in terms of the 2-D framework. The 2-D framework doesn't seem to change things too much. We can try to characterize a concept's PI in terms of the PI or the SI of others concepts, and we can try to characterize a concept's SI in terms of the PI or the SI of others' concepts, and we'll succeed to differing extents in different cases.

>Chalmers says that 'we should not mistake the linguistic expression [of
>1-intensions] for the real thing' (p. 13) Fair enough, but if there is a
>real thing here, should we be able to *unambiguously* refer to it. And it
>looks like we can't, for every attempt that one makes to refer to notional
>or relational content *via words* (and what else can one use?) can be
>interpreted in one of two ways, either 1-intensionally or 2-intensionally.
>(Actually, I think that there's more than one problem contained in the
>above, but it's lunchtime and I'm not going to rewrite it.)

Two things to say here. (1) If we characterize a PI or SI in terms of an expression E such that E's PI and SI are more or less the same, this problem won't really come up. (2) It seems to me that we can avoid ambiguity in any case simply by saying that we are talking about the PI or E, or the SI of E. So we can say: the PI of concept C corresponds more or less to the PI of E, where hopefully E has a more transparent structure. And same for the SI. So hopefully this gives us some sort of effability. Here I am just repeating what Brad said, I think.

Sometimes people say narrow content is ineffable basically because it is hard to characterize the PI of a concept C in terms of the *SI* of another concept E. (E.g., the narrow content of "water" is the function that picks out such-and-such in all counterfactual worlds.) But as I say in the paper, that seems to be something of an unfair requirement. Just as we can characterize the SI of one concept in

terms of the SI of others, we can characterize the PI of one concept in terms of the PI of others.

Re Kripke on rule-following, Tim wrote:

>(3) I'm worried about the implications Kripke's rule-following has for the
>accessability of notional content. Chalmers says that notional content is
>always (in principle, at least) accessible. It is internal to the thinker.
>But consider Kripke's plus/quus concerns. It's not exactly clear what
>these are, but here's a stab: Do I know what I mean by 'plus' merely by
>introspecting? No. In fact, I only mean something by plus by 'how I go
>on', that is, by the rule I follow **over time**. So how can I know what I
>mean by 'plus' **now** if what I do in fact mean is in part fixed by the
>rule that I follow? In a nutshell, the worry is that (certain) notional
>contents are fixed by rules that I follow over time, and these may be
>cognitively inaccessible to me **at a time**.

Hmm, interesting. I certainly don't claim to have a full answer to the Kripke/Wittgenstein problem. But the sense in which I have access to a PI is not (or at least not obviously) the sense in which I can specify a PI all at once, or grasp it all at one time, or some such. Think again of the case of "knowledge". The sense in which we have access to a PI is the sense in which, **given** a description of a scenario (or ultimately a world), we can say how and whether the PI applies to it.

It's best ultimately to do this with thoughts rather than concepts (because of problems with inscrutability of reference). The PI of a thought is a function from worlds to truth-values; we have access to the PI because given a qualitative description of a world, we can determine whether the PI of the thought is true. That's to say, we can evaluate the function in question at any world, and we can do this *a priori*.

It seems to me that we have something like this in the "plus" case. For any given statement involving plus -- " $56 + 65 = 121$ ", for example -- we're in a position to know whether it is true or false (of course if it is true in one world it is true in any world). So we have access to the PI of these statements.

It's true that in evaluating these statements, one is doing something new. And there is a deep philosophical puzzle about just what in my prior state determines that my answer to this question is "true" rather than "false". But I don't think we need to solve this puzzle here. We can just taken it as an intuitive datum there **is** a correct answer to give to these questions, upon rational reflection, and define the PI in terms of that.

The philosophical problem of just how this determinate intensionality is grounded remains, as it does for any semantic theory, but we don't need to solve all our problems at once! For my part, I think this is part of the project of "naturalizing intentionality", one that I haven't talked much about in this seminar. Specifically, what "natural" facts makes it the case that a thinker has a concept with one PI rather than another? Personally I think this will be grounded in a combination of functioning and phenomenology, by virtue of certain idealization principles which are built into our intentional concepts, but that's a long story.

I suppose another way the K/W problem comes up is in yielding a determinate PI for "plus" thoughts involving billion-digit numbers,

say, such that I will never in fact be able to give a determinate judgment. Here I need again to appeal to an idealization, going with what a less limited reasoner would say, or what would be the rationally correct thing to say. Again, the question of just what grounds that is tricky, but here I am simply appealing to the very strong intuition that there is indeed a rationally correct answer to these questions. If that intuition turned out to be wrong (perish the thought!), it might then turn out that the PIs of our concepts are less determinate than we think they are.

--Dave.

From owner-modality@LISTSERV.ARIZONA.EDU Tue Apr 27 11:49:56 1999

Date: Tue, 27 Apr 1999 11:48:10 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Erik A Herman <erikh@U.ARIZONA.EDU>

Subject: Intuitions on Components of Content

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: R

I would like to resolve my intuitions with the 2-Framework and I'm not quite sure how to do it. I've listed my intuitions on each of the 6 cases:

1) The twins relation to their environment is identical.

2) There is a difference between saying "Hesperus is Hesperus", and "Hesperus is Phosphorus". The first is just a figurative way of saying "a thing is itself", which HAS meaning, just not a very important one because it's so trivial and obvious. The second, on the other hand, is more like saying "two things are the same when you think about it". Content doesn't just denote something's physical embodiment but it's entire meaning. "the former is trivial and the latter is not"

3) Kripke's puzzle happens in real life. And the language issue doesn't do anything but explicate that Pierre doesn't see that THAT'S the "London" (however you wanna say it) that he's been told is pretty. Consider this real life scenario: I was told in phone conversations and e-mails "Mike is a cool guy." Then one night I was at a party and I met a guy, his name was Mike. Later, my friend asked if I had met Mike at the party, and being the flaky guy I am, said, "Well yes, I met a guy named Mike, why?" and she responded "THAT'S MIKE!" (meaning, that's THE Mike she had been speaking of in our correspondence). At which time I "linked/combined" these two, otherwise separated references. Now there doesn't seem to be anything confusing about this, and yet it is similar, in the relevant way, to the Londres case. If I had thought Mike was a jerk when I met him, this contradiction would be resolved in the "linking/combining" stage--analyzing my reasons for thinking he's a jerk with her reasons for thinking he's cool and weighing everything in the mix. I might even keep BOTH contents because they differ in their relation to Mike. To make the analogy hold, we could say that at the party he introduced himself as "Michel" or something stupid like that but this doesn't do any work.

4) My favorite video game is a car race (I forget what it's called) that is at the Flying J truckstop halfway to Phoenix (that's relevant). When I play it and I believe that "I'm" going to crash into a guardrail, I take evasive action. After I'm done playing, I get into the real car and I get yelled at by my girlfriend because I'm apparently driving like I'm still in the video game. The distinction is made in the differing degrees of indistructability. There is no difference IN KIND between the case where I am indestructible and the case where I might actually get hurt, just a

severe difference in degree of personal meaning they have. The same concept holds for the difference in degrees of recklessness in driving a car -vs- a motorcycle.

5) "Clark Kent" and "Superman" have two totally different meanings, the only similarity is that they occupy the same body.

6) This one has me stumped.

-erik h.

From owner-modality@LISTSERV.ARIZONA.EDU Mon May 3 13:05:52 1999

X-Accept-Language: en

Date: Mon, 3 May 1999 12:22:41 -0700

Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>

From: Josh Cowley <josh@MATH.ARIZONA.EDU>

Subject: PIs as narrow content

To: MODALITY@LISTSERV.ARIZONA.EDU

Status: R

I'm having a worry about PIs being something that is entirely in the head (i.e.. internal). In particular I'm not sure if they can do the work

we want them to if they are restricted to the head. Here is a sketch of the argument.

PIs are supposed to be functions from worlds to referents. Dave suggested that PIs might supervene on the physical/phenominal state of the individual. The problem is combining 2D semantics with a reductive theory semantics. A reductive theory of semantics is one which does not use any semantic terms in its terminology. Some of what people have been worried about in cashing out narrow content is how something that is *just* internal is going to pick out referents in the world. For example, if you are not allowed to use semantic terms it is hard to see how my concept "water" picks out water in the actual world. Since we are not allowed to use any sort of physical or causal chain to connect "water" to water, one wonders why it doesn't pick out toothpicks.

People have attempted to give reductive semantic theories which are in the head. I want to get on to why this is a problem for 2D semantics so I'll just give a quick example to motivate the problem. Conceptual role semantics (CR) claims that a concept gets its meaning by the role that it plays in a persons conceptual structure. For example, the role it plays in inferences, belief generation etc. Now this is clearly all in the head, but the question is whether it is enough to get reference. There are two ways we can go about getting the referent of "water." The first is to show the relationship between water and other concepts. For instance, "water" is a clear liquid which quenches thirst. It isn't an animal. It makes up the bulk of the lakes and rivers etc. This fits well with 2D semantics, but it defines "water" by other terms whose meaning is fixed. So this isn't really a reductive theory. The second way is to look at all the relations between all concepts, then find a set of objects and relations in the world which has a 1to1 correspondence to the relations of the concepts. The problem with this method is that you can always find some set of relations holding between objects which will be in 1to1 correspondence with the conceptual role relations.

That was a little sloppy, but others have argued for it more carefully and similar arguments have been raised against other internal

semantic theories. So suppose it is true that no strictly internal theory can determine a referent. Why does this cause problems for saying that PIs supervene on the physical state of the brain. Well, a PI is a function from worlds to referents. In order to be a function it has to pick out only one referent per world. But internal theories don't seem able to do this.

From owner-modality@LISTSERV.ARIZONA.EDU Mon May 3 13:25:00 1999
Date: Mon, 3 May 1999 13:23:05 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Re: PIs as narrow content
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

I may be missing Josh's point here, but I gather that he thinks that PI's are problematic because they are functions, and functions have to pick out unique referents, but PI's are committed to internal semantic theories, and these don't seem capable of pin-pointing unique referents.

I have a couple of comments, although I'm not entirely sure they are on track with what Josh is getting at. First, on the 2D view, both primary and secondary intensions are functions, and hence both are committed to returning unique referents as outputs. The problem, I suppose, is that only primary intensions are supposed to be confined to the head, and so are vulnerable to problems with internal theory semantics. It seems like one could argue that there isn't much of a distinction here however, and that primary and secondary intensions are functions that are either both both vulnerable to problems with internal semantics, or both okay, for roughly the same reasons. However we pick out water, as the clear drinkable liquid of a PI, or the H2O or XYZ of a SI, it seems that the story of how we manage to refer to what we do is going to pretty similar in both cases. So, we might have a problem with our semantics getting the unique referent of a function, but it will apply equally to watery stuff or H2O.

More generally, I think that worrying about reference re functions is an interesting problem, but it probably applies to some pretty standard thought (eg arithmetic). I take it this is Kripke's point in his "Kripkenstein" argument. So, I suppose this means we up against a puzzling situation, but it's no problem for PI's alone, as Josh seems to suggest.

Erik

On Mon, 3 May 1999, Josh Cowley wrote:

> I'm having a worry about PIs being something that is entirely in the
> head (i.e.. internal). In particular I'm not sure if they can do the
> work
> we want them to if they are restricted to the head. Here is a sketch
> of the argument.
>
> PIs are supposed to be functions from worlds to referents. Dave
> suggested that PIs might supervene on the physical/phenominal state of
> the individual. The problem is combining 2D semantics with a
> reductive theory semantics. A reductive theory of semantics is one
> which does not use any semantic terms in its terminology. Some of
> what people have been worried about in cashing out narrow content is
> how something that is *just* internal is going to pick out referents

> in the world. For example, if you are not allowed to use semantic
> terms it is hard to see how my concept "water" picks out water in the
> actual world. Since we are not allowed to use any sort of physical or
> causal chain to connect "water" to water, one wonders why it doesn't
> pick out toothpicks.
>
> People have attempted to give reductive semantic theories which are in
> the head. I want to get on to why this is a problem for 2D semantics
> so I'll just give a quick example to motivate the problem. Conceptual
> role semantics (CR) claims that a concept gets its meaning by the role
> that it plays in a persons conceptual structure. For example, the
> role it plays in inferences, belief generation etc. Now this is
> clearly all in the head, but the question is whether it is enough to
> get reference. There are two ways we can go about getting the
> referent of "water." The first is to show the relationship between
> water and other concepts. For instance, "water" is a clear liquid
> which quenches thirst. It isn't an animal. It makes up the bulk of
> the lakes and rivers etc. This fits well with 2D semantics, but it
> defines "water" by other terms whose meaning is fixed. So this
> isn't really a reductive theory. The second way is to look at all the
> relations between all concepts, then find a set of objects and
> relations in the world which has a 1to1 correspondence to the
> relations of the concepts. The problem with this method is that you
> can always find some set of relations holding between objects which
> will be in 1to1 correspondence with the conceptual role relations.
>
> That was a little sloppy, but others have argued for it more carefully
> and similar arguments have been raised against other internal
> semantic theories. So suppose it is true that no strictly internal
> theory can determine a referent. Why does this cause problems for
> saying that PIs supervene on the physical state of the brain. Well, a
> PI is a function from worlds to referents. In order to be a function
> it has to pick out only one referent per world. But internal theories
> don't seem able to do this.
>

Some advice for drivers on the University of Arizona campus:

"When passenger of foot heave in sight, tootle the horn. Trumpet
him melodiously at first, but if he still obstacles your passage
then tootle him with vigor."

--From a brochure of a car rental firm in Tokyo

Erik J Larson
erikl@U.Arizona.EDU

From owner-modality@LISTSERV.ARIZONA.EDU Tue May 4 10:47:25 1999
Date: Tue, 4 May 1999 10:43:42 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik A Herman <erikh@U.ARIZONA.EDU>
Subject: content
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

I think that only wide content exists. The subject's relation to a
concept is what the concept consists in. For instance, there is nothing
internal to Bert that distinguishes the two worlds. There is no "hard
content" where there is a one to one correspondence between the content

across worlds and the object itself. the mind just aims, shoots, and hopefully hits. It seems odd for me to be able to vary worlds and keep contents constant. Every subject is centered on their respective world--I don't get how the content could possibly hold across worlds or what such a concept could possibly mean. Where this is particularly clear to me is the case where Bert might say: "that is pretty" (and points). So I would guess that I agree with Loar if I understand his realization conditions correctly. In my view, contents are definitely context dependent and loose their meaning when extrapolated from their respective world.

Erik H.

From owner-modality@LISTSERV.ARIZONA.EDU Tue May 4 14:38:32 1999
Date: Tue, 4 May 1999 14:37:16 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony T Lane <atlane@U.ARIZONA.EDU>
Subject: Re: narrow content and justification
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Thony makes the case for the importance of narrow content and draws the following conclusion:

>
> I'm being unreasonable in the first case, and reasonable in the second.
> In an important sense, by belief that the thing is red is unjustified in
> the first case. If all that matters in beliefs is their wide content, we
> shouldn't have conflicting intuitions in these cases: there is no water
> that my beliefs are about. We clearly have conflicting intuitions about
> the cases, so there is some other kind of content that matters in
> justification.
>

I guess I have some sense of conflicting intuitions in this case, but it seems that the sense of conflict goes away if one considers exactly what question one is asking. If we ask whether the envatted individuals in 1 and 2 are justified in the beliefs they form, then clearly it seems that 1's belief is not justified and 2's is. But it seems to be a different question when we look at them from an external perspective and pronounce that, in both cases, their beliefs are unjustified.

It seems that the conflicting intuitions one has in this case just corresponds to which the different answers to the two questions. Ultimately, the two questions seems to correspond to PIs and SIs, and, once again, the disagreement arises from our different inclinations as to which is more essential.

Anthony

From owner-modality@LISTSERV.ARIZONA.EDU Mon May 3 13:29:50 1999
x-sender: agillies@pop.u.arizona.edu
Date: Mon, 3 May 1999 13:43:55 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: narrow content and justification
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Another reason why we need narrow content: although it looks like wide

content underwrites belief ascription, narrow content is what underwrites ascription of justification w.r.t. beliefs. Here's a sketch of how.

Consider the brain in a vat case. There is nothing in the world (outside the head) which justifies any of my beliefs, since my beliefs are about what I take to be the physical world made up of tables, chairs, water, books, cookies, and so on. But, by hypothesis, there are no such things outside my head; in reality it's just a big vat, with some liquid and electrodes. That's it. So my beliefs are not about (wide) e.g. water. But, borrowing a nice argument of Stew Cohen's, we still want to distinguish two different possible scenarios in the vat.

1. In the first scenario, suppose I have lots of beliefs about how water can refract light and how this influences the way things appear to me. So, I believe that I can't trust the shape of how things look to me when I believe they are partially immersed in water. Say I think that I see a bent stick, and draw the conclusion that it is bent. (Of course, the conclusion is false b/c I'm really just a brain in a vat.) Then suppose I also form the belief that the thing is partially immersed in water. Nevertheless (and without any other relevant beliefs) I continue to hold on to my belief that the thing is bent.

2. Now suppose everything is as in 1, except that upon forming the belief that thing is partially immersed in water, I withdraw my belief that the thing is bent.

I'm being unreasonable in the first case, and reasonable in the second. In an important sense, my belief that the thing is red is unjustified in the first case. If all that matters in beliefs is their wide content, we shouldn't have conflicting intuitions in these cases: there is no water that my beliefs are about. We clearly have conflicting intuitions about the cases, so there is some other kind of content that matters in justification.

Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Wed May 5 12:13:16 1999
X-Accept-Language: en
Date: Wed, 5 May 1999 12:09:28 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Josh Cowley <josh@MATH.ARIZONA.EDU>
Subject: More issues about a reductive theory
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

Some thoughts on yesterday's discussion. The PI of a concept is a function from worlds to referents. In order to give a reductive analysis of PIs we are going to have to find something internal that can act as a function from worlds to referents. It seems to me that nothing internal is going to really be a function from *worlds* to referents. Worlds are not psychological entities. Instead we are going to need some psychological entities which represent or perhaps are just in a 1to1 correspondence with worlds (for ease I'm going to suppose they represent worlds). So a reductive analysis of PIs is going to consist in finding a function from psychological representations to referents.

Now my worry is what the referents are going to be in this case. The referent either has to be something in the representation or something

in the world the representation is of. For example, consider the reductive primary intention of water given a representation of Twin Earth as its input. Either it has to pick out the representation of XYZ in my Twin Earth representation or it needs to pick XYZ in the Twin Earth world.

At the moment I don't have any argument that it is going to have to be one or the other. But I do have some general concerns. In order to make reference in the actual world work out, the range of the function better be things in worlds and not representations of them. If the range of the function is representations then in the actual world my term "Dave Chalmers" refers to that part of my actual world representation which represents Dave Chalmers. But we want "Dave Chalmers" to refer to Dave Chalmers himself, not a representation of him.

Now you might suggest that if you found a function from "Dave Chalmers" to the part of my world representation which represented Dave Chalmers, then you could simply extend the function by saying "Dave Chalmers" refers to whatever is represented by the part of my world representation which is picked out. But this ceases to be a reductive definition. Now we need to know what the Dave Chalmers part of the world representation represents. And that is just more semantics.

One last point; I think there are two assumptions underlying this worry. 1) PIs are internal (or supervene on what is in the head). 2) You (or your thought) doesn't have to be in a world in order to refer to something in the world. If you ditch both of these then you can use worlds themselves as the input to even a reductive function. Of course, this raises other problems.

Josh

From owner-modality@LISTSERV.ARIZONA.EDU Fri May 7 10:07:37 1999
x-sender: agillies@pop.u.arizona.edu
Date: Fri, 7 May 1999 10:21:52 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Anthony S Gillies <agillies@U.ARIZONA.EDU>
Subject: more on narrow content and justification
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

In looking over the Block & Stalnaker and Stalnaker pieces again, I'm getting a bit puzzled why people are so certain that there ain't any such beast as narrow content. This dovetails a bit with what I wrote last time on justification, so I'll expand on that as well.

Anthony responded to my call for narrow content this way:

>I guess I have some sense of conflicting intuitions in this case, but it
>seems that the sense of conflict goes away if one considers exactly what
>question one is asking. If we ask whether the envatted individuals in
>1
>and 2 are justified in the beliefs they form, then clearly it seems that
>1's belief is not justified and 2's is. But it seems to be a different
>question when we look at them from an external perspective and pronounce
>that, in both cases, their beliefs are unjustified.

I think that there is probably an ambiguity on 'justification' lurking around here. In one sense, S's belief in P is justified if S's belief in

P correlates or tracks truth to some specified degree. And whether a belief tracks truth is, presumably, a fact external to the believer. This, of course, is the sort of justification that Alvin wants to talk about. Anthony is right that in this sense of 'justification' the envatted believer has unjustified beliefs in both cases.

But Cohen's argument is that this is an uncomfortable consequence for externalists about justification. We have clear intuitions that go different ways in the two cases. And this is taken to suggest that there is another---perfectly respectable---notion of 'justification' which is essentially procedural in nature. Theories of justification are theories for rational belief acquisition, retention, and revision. And these are notions that can be explicated by appeal to stuff in the head (and nothing else). So there is an equally well-defined sense of 'justification' which relies on the mental states of the cognizer. And wide content cannot make sense out of this sort of justification. And that, in turn, suggests that there is another sort of content that must do the job.

It's been a long time coming, but I think people in epistemology are starting to come around to recognizing that there are at least these two different senses of 'justification', and that one need not be taken as more basic than the other. (Historical note: arguably, the internal sense of 'justification' is the sense we find in Descartes and Hume.) If that's right, then if it's also right that narrow content is needed for internalist epistemology, then we need not think that wide content is basic and narrow content derived, or vice-versa.

So, when Block&Stalnaker, and Stalnaker complain that narrow content can't be cleanly derived from wide content, an answer (probably an answer that Dave wants to give anyway) is basically a "So what?" reply: So narrow can't be derived from wide. Big deal. Internalist justification can't be derived from externalist justification, direct realism can't be derived from reliabilism. That just points to there being *two* phenomena in need of explanation. There is a clear sense of narrow content, and everybody knows it: it's that content that is needed (whatever the theory turns out to look like) that undergirds procedural justification. So there.

Cheers,
Thony

"Curious green ideas sleep furiously."

From owner-modality@LISTSERV.ARIZONA.EDU Wed May 5 22:40:17 1999
Date: Wed, 5 May 1999 22:38:57 -0700
Sender: "Philosophy 596B: Mind and Modality"
<MODALITY@LISTSERV.ARIZONA.EDU>
From: Erik J Larson <erikl@U.ARIZONA.EDU>
Subject: Frogs, flies, salamanders, and baseball hats
To: MODALITY@LISTSERV.ARIZONA.EDU
Status: R

On Wed, 5 May 1999, Josh Cowley wrote:

> Some thoughts on yesterdays discussion. The PI of a concept is a
> function from worlds to referents. In order to give a reductive
> analysis of PIs we are going to have to find something internal that
> can act as a function from worlds to referents. It seems to me that
> nothing internal is going to really be a function from *worlds* to
> referents. Worlds are not psychological entities. Instead we are
> going to need some psychological entities which represent or perhaps

> are just in a 1to1 correspondence with worlds (for ease I'm going to
> suppose they represent worlds). So a reductive analysis of PIs is
> going to consist in finding a function from psychological
> representations to referents.

My post starts here:

Some thoughts on primary intensions and reductive analysis. We want a reductive analysis of a primary intension that satisfies the following constraints. One, it is, as Josh says, "in the head" or psychological and two, it is a reduction of the primary intension as a function from centered worlds to referents. So, what is a psychological reduction of a function from centered worlds to referents? That's an interesting question. As Josh notes, "worlds" are not the sorts of objects that we can leave alone in a psychological reduction, because of course the brute notion of a possible world is not something that is likely to have a psychological realization simpliciter. So we need something more psychologically plausible, perhaps symbols in a LOT or at least something broadly representational. Now, I don't have any idea how this would work, and there is of course all the standard objections to reductive semantics of this sort. But it occurs to me that the characterization of a primary intension as a function is not problematic for psychological explanation per se, or if it is, it isn't any more puzzling than other functions that we employ in cognition. A rational agent can do arithmetic. That's in the head. It takes as input a specified domain--natural numbers, not worlds--and produces a unique output, the set of which constitute the range (a bunch more natural numbers in this case). But it seems that this is all in the head too. And so we have this miracle, internalist function--arithmetic! And it needs some psychological reduction, but we can't just take the NATURAL NUMBERS as psychologically brute, because those things are not psychological entities. And you can see the rest.

So, my point is just this. With Josh, I agree that there is a deep puzzle about how we could get an internal, psychological reduction of something like a primary intension, but unlike Josh (perhaps) I don't see this as a problem for a primary intension alone, especially when considered as a function whose domain and range are not prima facie psychological entities. This is just what we expect from the functions we commonly employ in cognition, whether balancing our check book or considering what the referent of a term in some world considered as actual is. Of course, there is a lot of ambiguity here about just what we mean by a reductive explanation in these cases, and just how far down it has to go. On a broad scale this sort of concern reflects the confusion over what to do with "rationality" and conditions of normativity generally in psychological reductions of "thinking".

So, consider:

1)we have the ability to use functions with potentially infinite domains and ranges in cognition and 2)by virtue of rational reflection, we have the ability to see whether the output of our functions is correct.

It seems to me that there is some puzzle about 1) (eg to what extent do we have to represent all this stuff cognitively, and how exactly is it represented?) but that 2) is the crux of the puzzle about giving a reductive analysis of a primary intension. I can return a referent as the output from a centered world. How is this referent the right one? Well, what is the intuitive, rational thing to say about cases where primary intensions are evaluated? That question is just the question of how we can get a psychological account of rationality, and that is a tricky one indeed. How do we get it from causal or functional processes, or even more puzzling, from a bunch of grey stuff in our skulls? So perhaps the

analysis of a primary intension is going to have to wait until we understand better how we do apriori reasoning, eg how we manage to be rational at all.

> ps Hey Rachael, did you get that paper turned in for Goldmann?

Erik L.

> Now my worry is what the referents are going to be in this case. The
> referent either has to be something in the representation or something
> in the world the representation is of. For example, consider the
> reductive primary intension of water given a representation of Twin
> Earth as its input. Either it has to pick out the representation of
> XYZ in my Twin Earth representation or it needs to pick XYZ in the
> Twin Earth world.

>
> At the moment I don't have any argument that it is going to have to be
> one or the other. But I do have some general concerns. In order to
> make reference in the actual world work out, the range of the function
> better be things in worlds and not representations of them. If the
> range of the function is representations then in the actual world my
> term "Dave Chalmers" refers to that part of my actual world
> representation which represents Dave Chalmers. But we want "Dave
> Chalmers" to refer to Dave Chalmers himself, not a representation of
> him.

>
> Now you might suggest that if you found a function from "Dave
> Chalmers" to the part of my world representation which represented
> Dave Chalmers, then you could simply extend the function by saying
> "Dave Chalmers" refers to whatever is represented by the part of my
> world representation which is picked out. But this ceases to be a
> reductive definition. Now we need to know what the Dave Chalmers part
> of the world representation represents. And that is just more
> semantics.

>
> One last point; I think there are two assumptions underlying this
> worry. 1) PIs are internal (or supervene on what is in the head). 2)
> You (or your thought) doesn't have to be in a world in order to refer
> to something in the world. If you ditch both of these then you can
> use worlds themselves as the input to even a reductive function. Of
> course, this raises other problems.

>
> Josh

Some advice for drivers on the University of Arizona campus:

"When passenger of foot heave in sight, tootle the horn. Trumpet
him melodiously at first, but if he still obstacles your passage
then tootle him with vigor."

--From a brochure of a car rental firm in Tokyo

Erik J Larson
erikl@U.Arizona.EDU

Part 2: Mental Content [869]

Part of [Contemporary Philosophy of Mind: An Annotated Bibliography](#).

Compiled by [David J. Chalmers](#), Department of Philosophy, University of Arizona, Tucson AZ 85721. E-mail: chalmers@arizona.edu.

- 2.1 [Propositional Attitudes and Folk Psychology](#) [285]
 - 2.1a [The Language of Thought](#) (Fodor) [54]
 - 2.1b [The Intentional Stance](#) (Dennett) [40]
 - 2.1c [Eliminativism](#) (Churchlands, etc) [57]
 - 2.1d [Propositional Attitudes, General](#) [42]
 - 2.1e [The Nature of Folk Psychology](#) [38]
 - 2.1f [The Simulation Theory](#) [54]
- 2.2 [Internalism and Externalism](#) [278]
 - 2.2a [Is Content in the Head?](#) (Putnam, Burge) [38]
 - 2.2b [Externalism and Psychological Explanation](#) [41]
 - 2.2c [Externalism and Mental Causation](#) [28]
 - 2.2d [Externalism and the Theory of Vision](#) [16]
 - 2.2e [Externalism and Computation](#) [11]
 - 2.2f [Externalism and Self-Knowledge](#) [72]
 - 2.2g [The Status of Narrow Content](#) [43]
 - 2.2h [Miscellaneous](#) [29]
- 2.3 [Theories of Content](#) [144]
 - 2.3a [Information-Based Accounts](#) (Dretske, etc) [29]
 - 2.3b [Asymmetric Dependence](#) (Fodor) [20]
 - 2.3c [Causal Accounts, General](#) [17]
 - 2.3d [Teleological Approaches](#) (Millikan, etc) [48]
 - 2.3e [Conceptual Role Approaches](#) [19]
 - 2.3f [Theories of Content, Misc](#) [11]
- 2.4 [Representation, General](#) [25]
- 2.5 [The Explanatory Role of Content](#) (Dretske, etc) [22]
- 2.6 [Concepts](#) [46]
- 2.7 [Meaning Holism](#) [26]
- 2.8 [Mental Content, Misc](#) [43]

2.1 The Status of Propositional Psychology

2.1a

The Language of Thought (Fodor)

Fodor, J. A. 1975. *The Language of Thought*. Harvard University Press.

Argues that thought involves computation upon representations, and that these are structured as sentences in a mental language. With linguistic and psychological evidence, and arguments that the mental language is innate.

Fodor, J. A. 1987. Why there still has to be a language of thought. In *Psychosemantics*. MIT Press.

Because it fits explanatory methodology, it coheres with the usual ontology of psychological processes, and it explains systematicity.

Fodor, J. A. 1978. Propositional attitudes. *Monist* 61:501-23. Reprinted in *RePresentations* (MIT Press, 1980).

About what PA's are, and why they're at the foundations of thought.

Fodor, J. 2001. Language, thought and compositionality. *Mind and Language* 16:1-15.

Abbott, B. 1995. Natural language and thought: Thinking in English. *Behavior and Philosophy* 23:49-55.

Bonjour, L. 1991. Is thought a symbolic process? *Synthese* 89:331-52.

Argues that symbol processing can't account for the intrinsically contentful nature of thought: using a symbol doesn't give understanding of its content. With defense against arguments from twin earth and conceptual-role semantics.

Braddon-Mitchell, D. & Fitzpatrick, J. 1990. Explanation and the language of thought. *Synthese* 83:3-29.

No need to postulate LOT: diachronic explanation is as good as synchronic, and high-level laws can exist without high-level causal connections.

Clapin, H. 1997. Problems with principle P. *Pacific Philosophical Quarterly* 78:261-??.

Clark, A. 1988. Thoughts, sentences and cognitive science. *Philosophical Psychology* 1:263-78.

Crane, T. 1990. The language of thought: No syntax without semantics. *Mind and Language* 5:187-213.

Davies, M. 1992. Aunty's own argument for the language of thought. In (J. Ezquerro & J. Larrazabal, eds)

Cognition, Semantics and Philosophy. Kluwer.

Dennett, D. C. 1977. A cure for the common code. *Mind*. Reprinted in *Brainstorms* (MIT Press, 1978).

Review of Fodor's LOT. Fodor's view is too strong: function, not structure, is criterial for content. The structure of a predictive theory need not be directly reflected in inner processing.

Dennett, D. C. 1975. Brain writing and mind reading. *Minnesota Studies in the Philosophy of Science* 7:403-15. Reprinted in *Brainstorms* (MIT Press, 1978).

On the explicit representation of belief: criteria, plausibility, and relationship to verbal reports and conscious judgments.

Dennett, D. C. 1990. Granny's campaign for safe science. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

A general treatment of Fodor, identifying him as arch-conservative mentalist.

DeWitt, R. 1995. Vagueness, semantics, and the language of thought. *Psyche* 1.

Dunlop, G. 1990. Conceptual dependency as the language of thought. *Synthese* 82:275-96.

Relates Schank's conceptual dependency to Fodor's LOT.

Egan, M. F. 1991. Propositional attitudes and the language of thought. *Canadian Journal of Philosophy* 21:379-88.

Contra two of Fodor's arguments for LOT. Complex causes need not have LOT constituency structure; and evidence from psychological theory falls short.

Field, H. 1978. Mental representation. *Erkenntnis* 13:9-18. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Analyzes belief into a relation between a person and an internal sentence, along with a semantic relation between that sentence and e.g. a proposition. With arguments against functionalist analyses, and against propositions.

Garson, J. W. 1997. Syntax in a dynamic brain. *Synthese* 110:343-55.

There are no good arguments for LOT of the form "The brain needs to do X, and X entails LOT". Considers X = concatenation, logical form, tracking, combinatorial encoding. Either LOT is weakened deeply or is unnecessary.

Garfield, J. 2000. Thought as language: A metaphor too far. *Protosociology* 14:85-101.

Gauker, C. 1995. *Thinking Out Loud: An Essay on the Relation between Thought and Language*. Princeton University Press.

Harman, G. 1973. *Thought*. Princeton University Press.

Harman, G. 1975. Language, thought, and communication. *Minnesota Studies in the Philosophy of Science* 7:270-298.

Argues that the primary role of language is in thought rather than in communication, and the language of thought incorporates natural language.

Harman, G. 1977. How to use propositions. *American Philosophical Quarterly*.

Harman, G. 1978. Is there mental representation? *Minnesota Studies in the Philosophy of Science* 9.

Hauser, L. 1995. Natural language and thought: Doing without mentalese. *Behavior and Philosophy* 23:41-47.

Heil, J. 1981. Does cognitive psychology rest on a mistake? *Mind* 90:321-42.

LOT confuses processes with descriptions of processes. Also, symbols cannot denote solely in virtue of structure, so must rely on human interpretation.

Kaye, L. J. 1994. The computational account of belief. *Erkenntnis* 40:137-53.

Kaye, L. J. 1995. The languages of thought. *Philosophy of Science* 62:92-110.

Knowles, J. 1998. The language of thought and natural language understanding. *Analysis* 58:264-272.

Loar, B. 1982. Must beliefs be sentences? *Philosophy of Science Association*.

Lycan, W. G. 1982. Toward a homuncular theory of believing. *Cognition and Brain Theory* 4:139-59.

Defends sententialism of the homuncular variety: little modules all the way in. Lots of pro-belief arguments.

Lycan, W. G. 1990. Mental content in linguistic form. *Philosophical Studies* 58:147-54.

Distinguishes varieties of Sententialism, reasonable vs. mad-dog.

Lycan, W. G. 1993. A deductive argument for the representational theory of thinking. *Mind and Language* 8:404-22.

Argues from the unboundedness of thinking and the need for a finite stock of elements to something like a language of thought. With remarks on connectionism and instrumentalism, and a reply by Stalnaker.

Laurence, S. & Margolis, E. 1997. Regress arguments for the language of thought. *Analysis* 57:60-66.

Marras, A. 1987. The weak and the strong representational theory of mind: Stich's interpretation of Fodor. *Dialogue* 26:349-55.

Matthews, R. J. 1989. The alleged evidence for representationalism. In (S. Silvers, ed) *Rerepresentation*. Kluwer.

Argues that contrary to some claims, cognitive psychology does not provide much support for a computational/representational theory of propositional attitudes. Specifically considers research in psycholinguistics and vision.

Matthews, R. J. 1991. Is there vindication through representationalism? In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Fodor's theory can't deal with inexplicit attitudes: the core/derivative distinction is untenable. But we can make sense of intentional causation and psychological explanation without explicit representation.

Millikan, R. G. 1993. On mentalese orthography. In (B. Dahlbom, ed) *Dennett and his Critics*. Blackwell.

On some problems typing tokens in the language of thought. There's no principled distinction between type-identical tokens and type-distinct tokens with an identity judgment. With interesting remarks on co-identification.

Pessin, A. 1995. Mentalese syntax: Between a rock and two hard places. *Philosophical Studies* 78:33-53.

Argues that there is no good way to individuate syntactic types in Mentalese. Neural typing, causal typing, and semantic typing all fail.

Rantala, V. & Vaden, T. 1997. Minds as connoting systems: Logic and the language of thought. *Erkenntnis* 46:315-334.

Rey, G. 1995. A not "merely empirical" argument for the language of thought. *Philosophical Perspectives* 9:201-22.

Schiffer, S. 1991. Does Mentalese have a compositional semantics? In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Argues that the language of thought need not have a compositional semantics; productivity and systematicity can be explained without it.

Schiffer, S. 1994. The language-of-thought relation and its implications. *Philosophical Studies* 76:263-85.

Schwartz, G. 1996. Symbols and thought. *Synthese* 106:399-407.

Sher, G. 1975. Sentences in the brain. *Philosophy and Phenomenological Research* 36:94-99.

On Danto's suggestion that beliefs are like sentences. Conventionality poses problems, as does differentiating between different sorts of attitudes.

Stalnaker, R. C. 1990. Mental content and linguistic form. *Philosophical Studies* 58:129-46.

Sterelny, K. 1983. Mental representation: What language is Brainese? *Philosophical Studies*, 43:365-82.

Motivates LOT and defends it against various objections: e.g. tacit belief, identity conditions, infinite regress, and semantic nativism.

Stich, S. P. 1978. Beliefs and subdoxastic states. *Philosophy of Science* 45:499-518.

Teng, N. Y. 1999. The language of thought and the embodied nature of language use. *Philosophical Studies* 94:237-251.

Tienson, J. 1990. Is this any way to be a realist? *Philosophical Psychology*.

Warmbrod, K. 1989. Beliefs and sentences in the head. *Synthese* 2:201-30.

Weller, C. 1997. Bonjour and mentalese. *Synthese* 113:251-63.

Yagisawa, T. 1994. Thinking in neurons: Comments on Stephen Schiffer's "The language-of-thought relation and its implications". *Philosophical Studies* 76:287-96.

2.1b The Intentional Stance (Dennett)

Dennett, D. C. 1978. *Brainstorms*. MIT Press.

Dennett, D. C. 1971. Intentional systems. *Journal of Philosophy* 68:87-106 Reprinted in *Brainstorms* (MIT Press, 1978).

Can view systems from physical stance, design stance, or intentional stance. Beliefs/desires are attributed under the intentional stance, with help from certain idealized norms of rationality and accuracy licensed by

evolution.

Dennett, D. C. 1981. Making sense of ourselves. *Philosophical Topics* 12:63-81. Reprinted in *The Intentional Stance* (MIT Press, 1987).

Reply to Stich 1981. Irrationality is misdesign (take design stance). Etc.

Dennett, D. C. 1987. *The Intentional Stance*. MIT Press.

Beliefs/desires are useful predictive attributions. This isn't inconsistent with a certain degree of realism (abstracta/illata distinction).

Dennett, D. C. 1988. *Precis of The Intentional Stance*. Behavioral and Brain Sciences.

TIS, with commentaries and replies.

Dennett, D. C. 1990. The interpretation of texts, people and other artifacts. *Philosophy and Phenomenological Research* (Supplement) 50.

Mental states are underdetermined: like interpreting a text, or finding an object's function. Even adaptationist teleology gives no fact of the matter.

Dennett, D. C. 1991. Real patterns. *Journal of Philosophy* 88:27-51.

Proposition attitudes have the ontological status of a noisy pattern that helps make sense of behavior. This degree of realism falls on a scale: Fodor > Davidson > Dennett > Rorty > Churchland.

Baker, L. R. 1987. Instrumentalism: Back from the brink? In *Saving Belief*. Princeton University Press.

Dennett vacillates between stance-dependence, -independence; e.g. on rationality, design features. Instrumentalism can't be rescued.

Baker, L. R. 1989. Instrumental intentionality. *Philosophy of Science* 56:303-16.

Bechtel, W. 1985. Realism, instrumentalism, and the intentional stance. *Cognitive Science* 9:265-92.

Dennett should be a realist, of the relative-to-environment variety.

Byrne, A. 1998. Interpretivism. *European Review of Philosophy* 3.

Cam, P. 1984. Dennett on intelligent storage. *Philosophy and Phenomenological Research* 45:247-62.

Clark, A. 1990. Belief, opinion and consciousness. *Philosophical Psychology*.

Argues contra Dennett and Smolensky that language is fundamental, not just an add-on.

Cohen, B. 1995. Patterns lost: Indeterminism and Dennett's realism about beliefs. *Pacific Philosophical Quarterly* 76:17-31.

Cummins, R. 1981. What can be learned from *Brainstorms*? *Philosophical Topics* 12:83-92.

Questioning Dennett on the bridge between intentional characterization and functional characterization.
Arguing for the importance of context.

Davies, D. 1995. Dennett's stance on intentional realism. *Southern Journal of Philosophy* 33:299-312.

Fodor, J. A. 1981. Three cheers for propositional attitudes. In *Representations*. MIT Press.

Dennett's rationality/intentional idealization assumptions should not be viewed as Platonic but epistemic.
PA's are real and play real roles.

Fodor, J. A. & LePore, E. 1993. Is intentional ascription intrinsically normative? In (B. Dahlbom, ed) *Dennett and His Critics*. Blackwell.

Against "interpretivism" about intentionality: projectivism is hopeless, and Dennett's arguments for
normativism (via charity and evolution) go wrong or beg the question.

Foss, J. 1994. On the evolution of intentionality as seen from the intentional stance. *Inquiry* 37:287-310.

Gauker, C. 1988. Objective interpretationism. *Pacific Philosophical Quarterly* 69:136-51.

Haugeland, J. 1993. Pattern and being. In (B. Dahlbom, ed) *Dennett and His Critics*. Blackwell.

Heitner, R. 2000. Is design relative or real? Dennett on intentional relativism and physical realism. *Minds and Machines* 10:267-83.

Hornsby, J. 1992. Physics, biology, and common-sense psychology. In (D. Charles & K. Lennon, eds) *Reduction, Explanation and Realism*. Oxford University Press.

Kukla, R. 2000. How to get an interpretivist committed. *Protosociology* 14:180-221.

Lyons, W. 1990. Intentionality and modern philosophical psychology, I. The modern reduction of
intentionality. *Philosophical Psychology* 3:247-69.

McLaughlin, B. & O'Leary-Hawthorne, J. 1995. Dennett's logical behaviorism. *Philosophical Topics*

22:189-258.

McLaughlin, B. 2000. Why intentional systems theory cannot reconcile physicalism with realism about belief and desire. *Protosociology* 14:145-157.

McCulloch, G. 1990. Dennett's little grains of salt. *Philosophical Quarterly* 40:1-12.

Dennett must be one of: realist, eliminativist, instrumentalist.

Narayanan, A. 1996. The intentional stance and the imitation game. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Nelkin, N. 1993. Patterns. *Mind and Language* 9:56-87.

Dennett's instrumentalism can't explain the acquisition of intentional concepts. Proposition attitudes are directly introspectible entities, although still theoretical and still patterns.

Price, H. 1995. Psychology in perspective. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind*. Kluwer.

Radner, D. & Radner, M. 1995. Cognition, natural selection, and the intentional stance. *International Studies in the Philosophy of Science* 9:109-19.

Richard, M. 1995. What isn't a belief? *Philosophical Topics* 22:291-318.

Richardson, R. C. 1980. Intentional realism or intentional instrumentalism? *Cognition and Brain Theory* 3:125-35.

Sharpe, R. 1989. Dennett's journey towards panpsychism. *Inquiry* 32:233-40.

Slors, M. 1996. Why Dennett cannot explain what it is to adopt the intentional stance. *Philosophical Quarterly* 46:93-98.

Stich, S. P. 1980. Headaches. *Philosophical Books* 21:65-73.

Critical review of *Brainstorms*, with response.

Stich, S. P. 1981. Dennett on intentional systems. *Philosophical Topics* 12:39-62. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Dennett has problems with rationality, realism, etc. Hard line/soft line: either intentional stance is too close to FP or too far away.

Webb, S. 1994. Witnessed behavior and Dennett's intentional stance. *Philosophical Topics* 22:457-70.

Yu, P. & Fuller, G. 1986. A critique of Dennett. *Synthese* 66:453-76.

Very thorough account of the evolution of Dennett's views. Elucidates abstracta/illata, criticizes intentional subpersonal psychology.

2.1c Eliminativism (Churchlands) [see also 4.3c]

Churchland, P. S. 1980. Language, thought, and information processing. *Nous* 14:147-70.

Sentential processing is out. Against Harman's mental English and Fodor's Mentalese. Arguments from learning, evolution, neuroscience, mental images.

Churchland, P. M. 1981. Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78:67-90. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Eliminate beliefs/desires, remnants of a stagnant folk theory.

Churchland, P. M. & Churchland, P. S. 1983. Stalking the wild epistemic engine. *Nous* 17:5-20. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

How to dethrone language and still handle content.

Churchland, P. M. 1985. On the speculative nature of our self-conception. *Canadian Journal of Philosophy Supplement* 11:157-173.

Reply to Foss 1985: EM is plausible, though certainly not applicable everywhere -- e.g. sensations will be reduced, not eliminated.

Churchland, P. M. 1989. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT Press.

14 glimpses of the neurophilosophical golden age.

Churchland, P. M. 1993. Theory, taxonomy, and methodology: A reply to Haldane's "Understanding folk". *Proceedings of the Aristotelian Society* 67:313-19.

Reply to Haldane 1988. Even observations can be reconceived. With remarks perceptual plasticity and propositions, and a rejoinder by Haldane.

Churchland, P. M. 1993. Evaluating our self-conception. *Mind and Language* 8:211-22.

It's "bad faith" to accept modern epistemology but to deny the possibility of eliminativism. On various objections: "functional kinds", "self-defeating", "what could falsify it?", "different purposes", "no alternatives".

Baker, L. R. 1987. The threat of cognitive suicide. In *Saving Belief*. Princeton University Press.

Elaborating the paradoxes of disbelieving in belief. Rational acceptability, assertion, and truth are all at risk.

Baker, L. R. 1988. Cognitive suicide. In (R. Grimm & D. Merrill, eds) *Contents of Thought*. University of Arizona Press.

Eliminativism is pragmatically incoherent, as it implies that language isn't meaningful and that the thesis isn't formulable. Folk psychology needn't be scientifically reduced to be true. With comments by Chastain, and reply.

Bertolet, R. 1994. Saving eliminativism. *Philosophical Psychology* 7:87-100.

Against Baker's cognitive-suicide arguments against eliminativism. We don't know what a replacement theory will look like, but that doesn't show that none is forthcoming.

Bickle, J. 1992. Revisionary physicalism. *Biology and Philosophy* 7:411-30.

Argues for a revisionary reduction of the propositional attitudes, rather than elimination or smooth reduction. Sentential aspects will go, but coarse-grained functional profiles and content will remain.

Blunt, P. K. 1992. A defense of folk psychology. *International Philosophical Quarterly* 32:487-98.

Boghossian, P. 1990. The status of content. *Philosophical Review* 99:157-84.

Irrrealism about mental content (and therefore truth-conditions) can't be made sense of. An error thesis presupposes factual truth-conditions, and a non-factualist thesis presupposes a non-deflationary theory of truth.

Boghossian, P. 1991. The status of content revisited. *Pacific Philosophical Quarterly* 71:264-78.

Reply to Devitt 1990.

Chater, N. & Oaksford, M. 1996. The falsity of folk theories: Implications for psychology and philosophy. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Clark, A. 1996. Dealing in futures: Folk psychology and the role of representations in cognitive science. In (R. McCauley, ed) *The Churchlands and their Critics*. Blackwell.

Cling, A. 1989. Eliminative materialism and self-referential inconsistency. *Philosophical Studies* 56:53-75.

Unbelief in belief is not incoherent. Argues with Baker.

Cling, A. 1990. Disappearance and knowledge. *Philosophy of Science* 57:226-47.

Cling, A. 1991. The empirical virtues of belief. *Philosophical Psychology* 4:303-23.

Cognitive states like belief are necessary to explain the dependence of behavior on perceptual features of the environment. Informational states alone are not enough, as they can't explain selective response to features.

Devitt, M. 1990. Transcendentalism about content. *Pacific Philosophical Quarterly* 71:247-63.

Against Boghossian's critique: the eliminativism will express her claim in a new framework, so appeals to truth beg the question. With a response.

Devitt, M. & Rey, G. 1991. Transcending transcendentalism. *Pacific Philosophical Quarterly* 72:87-100.

Rejoinder to Boghossian 1990.

Foss, J. E. 1985. A materialist's misgivings about eliminative materialism. *Canadian Journal of Philosophy Supplement* 11:105-33.

EM needs much more evidence before being so gung ho.

Graham, G. & Horgan, T. 1992. Southern fundamentalism and the end of philosophy. In (E. Villanueva, ed) *Truth and Rationality*. Ridgeview.

Greenwood, J. D. 1991. Reasons to believe. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

Argues that folk psychological states exist, even if they aren't useful as causal explanation. We have independent reason to believe in them, e.g. from self-knowledge. They're useful in social psychology, too.

Greenwood, J. D. 1992. Against eliminative materialism: from folk psychology to Volkerpsychologie. *Philosophical Psychology* 5:349-68.

Haldane, J. 1988. Understanding folk. *Aristotelian Society Supplement* 62:222-46.

Argues that folk psychology is not a theory, and that psychological knowledge is a pre-theoretical given.

With remarks on laws, the prediction of behavior, and neuroscience.

Hannan, B. 1990. 'Non-scientific realism' about propositional attitudes as a response to eliminativist arguments. *Behavior and Philosophy* 18:21-31.

Hannan, B. 1993. Don't stop believing: the case against eliminative materialism. *Mind and Language* 8:165-179.

A bundle of arguments against eliminativism, e.g. from incoherence, the lack of alternatives, and against the folk-theory-theory. With commentary.

Horgan, T. & Woodward, J. 1985. Folk psychology is here to stay. *Philosophical Review* 94:197-225. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Defending folk psychology against the arguments of Churchland and Stich: e.g. incompleteness, stagnation, irreducibility, dual-control, modularity, and unfalsifiability. Even with no neat reduction, folk psychology may be OK.

Horgan, T. & Graham, G. 1990. In defense of Southern Fundamentalism. *Philosophical Studies* 62:107-134.

FP is almost certainly true, irrespective of scientific absorbability or the language of thought. FP's commitments are austere, and mostly behavioral. Arguments from semantic competence and conceptual conservatism.

Horgan, T. 1993. The austere ideology of folk psychology. *Mind and Language*.

Argues that FP is not committed to much. The austere conception is supported by intuitions, conservatism, and the inconceivability of dropping it. Responds to phlogiston objections: they are not analogous.

Horst, S. 1995. Eliminativism and the ambiguity of 'belief'. *Synthese* 104:123-45.

Clarifies different senses of "theoretical" and "belief". Some beliefs are relevantly theoretical (dispositional, infra-conscious, unconscious ones), but conscious occurrent beliefs are not, and so can't be eliminated.

Jackson, F. & Pettit, P. 1990. In defense of folk psychology. *Philosophical Studies* 59:31-54.

FP holds that beliefs/desires play a certain functional role, and it's almost certain that objects playing that role exist, so FP is fine, whether or not propositional attitudes are good scientific entities.

Jacoby, H. 1985. Eliminativism, meaning and qualitative states. *Philosophical Studies*.

Even if nothing satisfies all or most common-sense properties of mental terms, reference can still be fixed

under a Putnam style theory of meaning. (More about qualia than about intentional states.)

Kitcher, P. S. 1984. In defense of intentional psychology. *Journal of Philosophy* 81:89-106.

The Churchlands underestimate the resources of intentional psychology.

Lahav, R. 1992. The amazing predictive power of folk psychology. *Australasian Journal of Philosophy* 70:99-105.

Melnyk, A. 1996. Testament of a recovering eliminativist. *Philosophy of Science* 63:S185-93.

O'Brien, G. 1987. Eliminative materialism and our psychological self-knowledge. *Philosophical Studies* 52:49-70.

Uses empirical evidence to argue that there is prelinguistic awareness, so nominalistic arguments for eliminativism fail. And some awareness is innate, so we can't reconceive things in less than evolutionary time.

Ramsey, W. 1990. Where does the self-refutation objection take us? *Inquiry* 33:453-65.

The self-refutation objection reduces to other standard objections: counterexample, promissory note or reductio.

Ramsey, W. , Stich, S. P. & Garon, J. 1991. Connectionism, eliminativism, and the future of folk psychology. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

If connectionism is true, then eliminativism is true, as you can't isolate the causal role of individual beliefs in a connectionist system.

Reppert, V. 1991. Ramsey on eliminativism and self-refutation. *Inquiry* 34:499-508.

Response to Ramsey 1990: If there are no beliefs and so no assertions, there is no identifiable propositional content, and truth and knowledge are out. Eliminativism is pragmatically self-refuting.

Reppert, V. 1992. Eliminative materialism, cognitive suicide, and begging the question. *Metaphilosophy* 23:378-92.

A careful analysis of whether self-refutation arguments against eliminativism beg the question by supposing that assertion requires belief. An account of what it is to beg the question, and a comparison to arguments about vitalism.

Resnick, P. 1994. Intentionality is phlogiston. In (E. Dietrich, ed) *Thinking Computers and Virtual*

Persons. Academic Press.

Richards, G. 1996. On the necessary survival of folk psychology. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Robinson, W. S. 1985. Toward eliminating Churchland's eliminationism. *Philosophical Topics* 13:60-67.

There's no reason to abandon FP, even if it doesn't reduce.

Rosenberg, A. 1991. How is eliminative materialism possible? In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.

Explaining how singular causal claims based on FP may be true even if FP is false; by analogy with phlogiston, and also because of near-vacuousness. EM isn't incoherent, as we can use a non-intentional replacement for belief.

Rosenberg, A. 1999. Naturalistic epistemology for eliminative materialists. *Philosophy and Phenomenological Research* 59:335-358.

Saidel, E. 1992. What price neurophilosophy? *Philosophy of Science Association* 1:461-68.

Folk psychology is compatible with neuroscientific models, but it need not smoothly reduce to neuroscience to have an important role.

Schouten, M. K. D. & de Jong, H. L. 1998. Defusing eliminative materialism: Reference and revision. *Philosophical Psychology* 11:489-509.

Schwartz, J. 1991. Reduction, elimination, and the mental. *Philosophy of Science* 58:203-20.

Stich, S. P. 1991. Do true believers exist? *Aristotelian Society Supplement* 65:229-44.

Eliminativism may have no determinate truth-conditions: if folk psychology is a poor theory, the question of whether or not "belief" refers may be empty.

Stich, S. P. 1992. What is a theory of mental representation? *Mind* 101:243-61.

Philosophical analysis isn't sufficient to understand intentional concepts; real cognitive science is required, with conceptual revision. The truth of eliminativism will be relative to the theory of reference that we choose.

Stich, S. P. 1996. Deconstructing the mind. In *Deconstructing the Mind*. Oxford University Press, 1996.

Taylor, K. A. 1994. How not to refute eliminative materialism. *Philosophical Psychology* 7:101-125.

Against transcendental arguments against eliminativism. These fail on their own terms, and even if successful they would not establish causal/explanatory relevance for the attitudes, which is the real key for folk psychology.

Tomberlin, J. 1994. Whither Southern Fundamentalism? In (E. Villanueva, ed) *Truth and Rationality*, Ridgeview.

Trout, J. D. 1991. Belief attribution in science: Folk psychology under theoretical stress. *Synthese* 87:379-400.

Wright, C. 1996. Can there be a rationally compelling argument for anti-realism about ordinary ("folk") psychology? In (E. Villanueva, ed) *Content*. Ridgeview.

2.1d Propositional Attitudes, General

Audi, R. 1994. Dispositional beliefs and dispositions to believe. *Nous* 28:419-34.

Baker, L. R. 1987. *Saving Belief*. Princeton University Press.

Beliefs are OK, despite no physicalist reduction of content.

Baker, L. R. 1993. What beliefs are not. In (S. Wagner & R. Warner, eds) *Naturalism: A Critical Appraisal*. University of Notre Dame Press.

Against beliefs construed as physically realized internal causes of behavior: syntax of these states can't be determinate, and their explanatory role wrt causation leads to a circle. Belief is irreducible.

Baker, L. R. 1994. Attitudes as nonentities. *Philosophical Studies* 76:175-203.

Balaguer, M. 1998. Attitudes without propositions. *Philosophy and phenomenological research* 58:805-26.

Bennett, J. 1991. Analysis without noise. In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.

Remarks on the conceptual analysis of belief/desire attribution. On the roles of causation, inner-route explanations, belief-desire-action triangles, teleology, unity, the presumption of simplicity, and evolution.

Bennett, J. 1991. Folk-psychological explanations. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

On requirements for belief/desire explanations: input/output patterns, the unity condition (i.e. no single

associated mechanism), and teleological bases for generalizations, e.g. through evolution or educability.

Ben-Yami, H. 1997. Against characterizing mental states as propositional attitudes. *Philosophical Quarterly* 186:84-89.

Butler, K. 1992. The physiology of desire. *Journal of Mind and Behavior* 13:69-88.

Argues that desire will smoothly reduce to a neurophysiological kind.

Clark, A. 1991. Radical ascent. *Aristotelian Society Supplement* 65:211-27.

The conditions on being a believer are mostly behavioral; to claim otherwise is to fall into a "modularity trap". A counterfactual account of mental causation is enough. With a defense of mentality for giant look-up tables.

Clark, A. 1994. Beliefs and desires incorporated. *Journal of Philosophy* 91:404-25.

Cohen, L. J. 1996. Does belief exist? In (A. Clark & P. Millican, eds) *Connectionism, Concepts, and Folk Psychology*. Oxford University Press.

Crimmins, M. 1992. Tacitness and virtual beliefs. *Mind and Language* 7:240-63.

Davies, D. 1995. Davidson, indeterminacy, and measurement. *Acta Analytica* 10:37-56.

Davies, D. 1998. On gauging attitudes. *Philosophical Studies* 90:129-54.

Egan, M. F. 1989. What's wrong with the Syntactic Theory of Mind. *Philosophy of Science* 56:664-74.

Stich is confused about type-token, syntax/content, etc.

Fodor, J. A. 1986. Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind* 94:76-100. Reprinted in *A Theory of Content and Other Essays* (MIT Press, 1990).

A taxonomy of positions on the representation of propositional attitudes: dividing up via questions about realism, functionalism, monadicity, and truth-conditions. With arguments for structured representations.

Frankish, K. 1998. A matter of opinion. *Philosophical Psychology* 11:423-442.

Garfield, J. 1988. *Belief in Psychology: A Study in the Ontology of Mind*. MIT Press.

Graham, G. & Horgan, T. 1988. How to be realistic about folk psychology. *Philosophical Psychology* 1.

Jacquette, D. 1990. Intentionality and Stich's theory of brain sentence syntax. *Philosophical Quarterly*, 40:169-82.

Things are only syntactic (in SS's sense) in virtue of intentionality. True.

Lycan, W. G. 1986. Tacit belief. In (R. Bogdan, ed) *Belief: Form, Content, and Function*. Oxford University Press.

Maloney, J. C. 1990. It's hard to believe. *Mind and Language* 5:122-48.

Manfredi, P. A. 1993. Tacit beliefs and other doxastic attitudes. *Philosophia*.

Argues that there are no tacit beliefs: dispositions to believe can do all the explanatory work at lower cost. With some remarks on subdoxastic states, and the difference between belief and opinion.

Matthews, R. J. 1994. The measure of mind. *Mind* 103:131-46.

A theory of propositional attitude ascription as like numerical measurement.

Millikan, R. G. 1986. Thoughts without laws: Cognitive science with content. *Philosophical Review* 95:47-80.

Folk psychology isn't a theory about laws, but about proper functions. desires are identified by proper functions; beliefs by Normal explanations.

Moser, P. K. 1990. Physicalism and intentional attitudes. *Behavior and Philosophy* 18:33-41.

Peacocke, C. 1983. Between instrumentalism and brain-writing. In *Sense and Content*. Oxford University Press.

Instrumentalism about belief can't be right, because of Martian marionettes, but the language of thought is too strong a requirement. A state's structured content may reside in its pattern of relations to other states.

Possin, K. 1986. The case against Stich's Syntactic Theory of Mind. *Philosophical Studies* 49:405-18.

Stich is wrong, circular, and representational anyway.

Pratt, I. 1993. Analysis and the attitudes. In (S. Wagner & R. Warner, eds) *Naturalism: A Critical Appraisal*. University of Notre Dame Press.

Pylyshyn, Z. W. 1987. What's in a mind? *Synthese* 70:97-122.

Must individuate mental states by semantics, not just by function.

- Recanati, F. 1997. Can we believe what we do not understand? *Mind and Language* 12:84-100.
- Robinson, W. S. 1990. States and beliefs. *Mind* 99:33-51.
- Schwartz, J. 1992. Propositional attitude psychology as an ideal type. *Topoi* 11:5-26.
- Smith, D. M. 1994. Toward a perspicuous characterization of intentional states. *Philosophical Studies* 74:103-20.
- Sobel, D. & Copp, D. 2001. Against direction of fit accounts of belief and desire. *Analysis* 61:44-53.
- Sperber, D. 1997. Intuitive and reflective beliefs. *Mind and Language* 12:67-83.
- Stich, S. P. 1983. *From Folk Psychology to Cognitive Science*. MIT Press.
- Beliefs/desires are out, new Syntactic Theory is in.
- Stich, S. P. 1984. Relativism, rationality, and the limits of intentional ascription. *Pacific Philosophical Quarterly*.
- Stone, T. & Young, A. W. 1997. Delusions and brain injury: The philosophy and psychology of belief. *Mind and Language* 12:327-364.
- Von Eckardt, B. & Poland, J. 2000. In defense of the standard view. *Protosociology* 14:312-331.
- Weatherall, P. 1996. What do propositions measure in folk psychology? *Philosophical Psychology* 9:365-80.

2.1e

The Nature of Folk Psychology

- Blackburn, S. 1992. Theory, observation, and drama. *Mind and Language* 7:187.
- Bogdan, R. G. (ed) 1991. *Mind and Common Sense: Philosophical Essays on Commonsense Psychology*. Cambridge University Press.
- Botterill, G. 1996. Folk psychology and theoretical status. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Churchland, P. M. 1988. Folk psychology and the explanation of human behavior. *Proceedings of the Aristotelian Society* 62:209-21. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Folk psychology is a theory: defense against objections from logicity, softness of laws, practical function, behavior, and simulation. It needn't be a deductive-nomological theory; e.g. it might be based on prototypes.

Clark, A. 1987. From folk psychology to naive psychology. *Cognitive Science* 11:139-54.

Folk psychology isn't all that bad. It survived evolution after all.

Collins, J. 2000. Theory of mind, logical form and eliminativism. *Philosophical Psychology* 13:465-490.

Dennett, D. C. 1991. Two contrasts: Folk craft vs folk science and belief vs opinion. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

FP is craft, not theory. Opinions rather than beliefs are interesting.

Falvey, K. 1999. A natural history of belief. *Pacific Philosophical Quarterly* 80:324-345.

Fletcher, G. 1995. *The Scientific Credibility of Folk Psychology*. Lawrence Erlbaum.

Fletcher, G. 1995. Two uses of folk psychology: Implications for psychological science. *Philosophical Psychology* 8:375-88.

Goldman, A. 1992. The psychology of folk psychology. *Behavioral and Brain Sciences*.

On the psychology of self-ascription of mental states. Functionalism has serious problems, as we don't have direct access to causal roles. Defends a qualia-based account, even for propositional attitudes.

Gopnik, A. 1990. Developing the idea of intentionality: Children's theories of mind. *Canadian Journal of Philosophy* 20:89-114.

On the development of folk-psychological concepts in children. First the appearance/reality distinction, then more complex theories of perception, representation, and belief. Implications for the status of folk psychology.

Gopnik, A. & Wellman, H. 1992. Why the child's theory of mind really is a theory. *Mind and Language* 7:145-71.

Graham, G. 1987. The origins of folk psychology. *Inquiry* 30:357-79.

Greenwood, J. D. (ed) 1991. *The Future of Folk Psychology: Intentionality and Cognitive Science*.

Cambridge University Press.

Leon, M. 1998. The unnaturalness of the mental: The status of folk psychology. *Southern Journal of Philosophy* 36:367-92.

Lycan, W. G. 1997. Folk psychology and its liabilities. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Margolis, J. 1991. The autonomy of folk psychology. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

McDonough, R. 1991. A culturalist account of folk psychology. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

Morton, A. 1980. *Frames of Mind*. Oxford University Press.

Morton, A. 1991. The inevitability of folk psychology. In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.

Morton, A. 1996. Folk psychology is not a predictive device. *Mind* 105:119-37.

Pettit, P. 2000. How the folk understand folk psychology. *Protosociology* 14:26-38.

Place, U. T. 1996. Folk psychology from the standpoint of conceptual analysis. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Pratt, I. 1996. Encoding psychological knowledge. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Preston, J. M. 1989. Folk psychology as theory or practice? The case for eliminative materialism. *Inquiry* 32:277-303.

Defending the claim that folk psychology is an empirical pre-scientific theory, with its own laws. In a particular, a detailed reply to the criticisms in Wilkes 1984.

Robinson, W. S. 1996. Mild realism, causation, and folk psychology. *Philosophical Psychology* 8:167-87.

Schwitzgebel, E. 2001. In-between believing. *Philosophical Quarterly* 51:76 - 82.

Sehon S. R. 1997. Natural kind terms and the status of folk psychology. *American Philosophical Quarterly* 34:333-44.

Sharpe, R. 1987. The very idea of a folk psychology. *Inquiry* 30:381-93.

Smith, B. C. 1996. Does science underwrite our folk psychology? In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Stemmer, N. 1995. A behaviorist account to theory and simulatyon theories of folk psychology. *Behavior and Philosophy* 23:29-41.

Sterelny, K. 1998. Intentional agency and the metarepresentation hypothesis. *Mind and Language* 13:11-28.

Stich, S. P. & Ravenscroft, R. 1994. What is folk psychology? *Cognition* 50:447-68. Reprinted in (Stich) *Deconstructing the Mind*. Oxford University Press, 1996.

Distinguishes internal and external accounts of folk psychology (mechanisms vs systematizations), and various versions of each of these. Only some are compatible with eliminativist arguments.

von Eckardt, B. 1997. The empirical naivete in the current philosophical conception of folk psychology. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Wilkes, K. V. 1984. Pragmatics in science and theory in common sense. *Inquiry* 27:339-61.

Wilkes, K. V. 1991. The relationship between scientific psychology and common-sense psychology. *Synthese* 89:15-39.

Common-sense psychology is no theory at all, and not in competition with scientific psychology. CSP is particular, rich, vague; SP is general, austere, precise. CSP will be neither subsumed nor eliminated by SP.

Wilkes, K. V. 1991. The long past and the short history. In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.

Argues that commonsense and scientific psychology are quite distinct in their aims, scope, framework, and nature, but have been confused by philosophy. With support from historical considerations.

2.1f

The Simulation Theory

Arkway, A. 2000. The simulation theory, the theory theory and folk psychological explanation. *Philosophical Studies* 98:115-137.

- Carruthers, P. 1996. Simulation and self-knowledge: A defence of the theory-theory. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.
- Carruthers, P. & Smith, P. 1996. *Theories of Theories of Mind*. Cambridge University Press.
- Cruz, J. L. H. 1998. Mindreading: Mental state ascription and cognitive architecture. *Mind and Language* 13:323-340.
- Currie, G. 1995. Visual imagery as the simulation of vision. *Mind and Language* 10:25-44.
- Currie, G. 1996. Simulation-theory, theory-theory, and the evidence from autism. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.
- Currie, G. & Ravenscroft, I. 1997. Mental simulation and motor imagery. *Philosophy of Science* 64:161-80/
- Currie, G. 1998. Pretence, pretending, and metarepresenting. *Mind and Language* 13:35-55.
- Davies, M. 1992. The mental simulation debate. In (E. Villanueva, ed) *Truth and Rationality*. Ridgeview.
- Davies, M. & Stone, T. (eds) 1995. *Folk Psychology: The Theory of Mind Debate*. Blackwell.
- Davies, M. & Stone, T. (eds) 1995. *Mental Simulation: Evaluations and Applications*. Blackwell.
- Freeman, N. H. 1995. Theories of mind in collision: Plausibility and authority. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.
- Fuller, G. 1995. Simulation and psychological concepts. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.
- Goldman, A. 1989. Interpretation psychologized. *Mind and Language* 4:161-85. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.
- Goldman, A. 1992. In defense of the simulation theory. *Mind and Language* 7:104-119. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.
- Goldman, A. 1996. Simulation and interpersonal utility. In (L. May, M. Friedman, & A. Clark, eds) *Mind and Morals: Essays on Ethics and Cognitive Science*. MIT Press.
- Goldman, A. 2000. Folk psychology and mental concepts. *Protosociology* 14:4-25.

Gopnik, A. & Wellman, H. M. 1995. Why the child's theory of mind really is a theory. *Mind and Language*. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Gopnik, A. and Meltzoff, AN. 1998. Theories vs. modules: To the max and beyond. A reply to Poulin-Dubois and to Stich and Nichols. *Mind and Language* 13:450-456.

Gordon, R. M. 1986. Folk psychology as simulation. *Mind and Language* 1:158-71. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

FP is a strategy for prediction via simulation; an ability, not a theory.

Gordon, R. M. 1992. The simulation theory: objections and misconceptions. *Mind and Language* 7:11-34. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Gordon, R. M. & Barker, J. A. 1994. Autism and the "theory of mind" debate. In (G. Graham & G. L. Stephens, eds) *Philosophical Psychopathology*. MIT Press.

Gordon, R. M. 1995. Simulation without introspection or inference from me to you. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Gordon, R. M. 1996. Sympathy, simulation, and the impartial spectator. In (L. May, M. Friedman, & A. Clark, eds) *Mind and Morals: Essays on Ethics and Cognitive Science*. MIT Press.

Gordon, R. M. 1996. 'Radical' simulationism. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Gordon, R. M. 2000. Sellars's Rylean ancestors revisited. *Protosociology* 14:102-114.

Greenwood, J. D. 1999. Simulation, theory-theory and cognitive penetration: No "instance of the fingerpost". *Mind and Language* 14:32-56.

Heal, J. 1986. Replication and functionalism. In (J. Butterfield, ed) *Language, Mind, and Logic*. Cambridge University Press. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Heal, J. 1994. Simulation vs. theory-theory: What is at issue? In (C. Peacocke, ed) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford University Press.

Heal, J. 1995. How to think about thinking. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Heal, J. 1996. Simulation and cognitive penetrability. *Mind and Language* 11:44-67.

Heal, J. 1996. Simulation, theory, and content. In (P. Carruthers & P. Smith, eds) *Theories of Theories of*

Mind. Cambridge University Press.

Heal, J. 1998. Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind and Language* 13:477-498.

Heal, J. 2000. Understanding other minds from the inside. *Protosociology* 14:39-55.

Henderson, D. 1996. Simulation theory versus theory theory: A difference without a difference in explanations. *Southern Journal of Philosophy* 34:65-93.

Kuhberger, A, Perner, J. , Schulte, M. , & Leingruber, R. 1995. Choice or no choice: Is the Langer effect evidence against simulation? *Mind and Language* 10:423-36.

Leslie, A. M. & German, T. P. 1995. Knowledge and ability in "theory of mind": A one-eyed overview of a debate. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Levin, J. 1995. Folk psychology and the simulationist challenge. *Acta Analytica* 10:77-100.

Nichols, S. , Stich, S. , & Leslie, A. 1995. Choice effects and the ineffectiveness of simulation. *Mind and Language* 10:437-45.

Nichols, S. , Stich, S. , Leslie, A. , Klein, D. 1996. Varieties of off-line simulation. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Nichols, S. & Stich, S. 1998. Rethinking co-cognition: A reply to Heal. *Mind and Language* 13:499-512.

Perner, J. 1994. The necessity and impossibility of simulation. In (C. Peacocke, ed) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford University Press.

Perner, J. 1996. Simulation as explication of predication-implicit knowledge about the mind: Arguments for a simulation-theory mix. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Perner, J. , Gschaidner, A. , Kuhberger, A. & Schrofner, S. 1999. Predicting others through simulation or by theory? A method to decide. *Mind and Language* 14:57-79.

Pust, J. 1999. External accounts of folk psychology, eliminativism, and the simulation theory. *Mind and Language* 14:113-130.

Ruffman, T. 1996. Do children understand the mind by means of a simulation or a theory? Evidence from their understanding of inference. *Mind and Language* 11:388-414.

Scholl, B. J. & Leslie, A. M. 1999. Modularity, development and "theory of mind." *Mind and Language* 14:131-153.

Schwitzgebel, E. 1999. Representation and desire: a philosophical error with consequences for theory-of-mind research. *Philosophical Psychology* 12:157-180.

Stich, S. P. & Nichols, S. 1993. Folk psychology: simulation or tacit theory? *Mind and Language* 7:35-71. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Stich, S. P. & Nichols, S. 1995. Second thoughts on simulation. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Stich, S. P. & Nichols, S. 1997. Cognitive penetrability, rationality, and restricted simulation. *Mind and Language* 12:297-326.

Stich, S. & Nichols, S. 1998. Theory theory to the max. *Mind and Language* 13:421-449.

Stone, T. & Davies, M. 1996. The mental simulation debate: A progress report. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Wilkerson, W. S. 2001. Simulation, theory, and the frame problem: the interpretive moment. *Philosophical Psychology* 14:141-153.

2.2

Internalism and Externalism [see also 1.5d]

2.2a Is Content in the Head? (Putnam, Burge)

Antony, M. 1993. Social relations and the individuation of thought. *Mind* 102:247-61.

Bilgrami, A. 1987. An externalist account of psychological content. *Philosophical Topics* 15:191-226.

Developing an externalist account consistent with psychological explanation. Contra Burge, social links aren't constitutive of content. Causal links are indirectly constitutive of content, via our conceptions.

Brueckner, A. 1995. The characteristic thesis of anti-individualism. *Analysis* 55:146-48.

Bruns, M. & Soldati, G. 1997. Object-dependent and property-dependent concepts. *Dialectica* 48:185-208.

Burge, T. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4:73-122.

Belief contents are not fully determined by internal state, as the linguistic community plays an important role: arthritis, brisket, contract, sofa, etc. So mental states are not individuated individualistically.

Burge, T. 1982. Other bodies. In (A. Woodfield, ed) *Thought and Object*. Oxford University Press.

On Putnam's Twin Earth. Natural kind terms are not indexical. Even de dicto attitudes are not in the head; they presuppose the existence of other things.

Burge, T. 1986. Intellectual norms and foundations of mind. *Journal of Philosophy* 83:697-720.

On non-individualist elements due to by intellectual norms in the community, to which meanings are answerable. Even meaning-giving truths can be doubted. With remarks on sofas/safos, and on linguistic meaning vs. cognitive value.

Butler, K. 1993. Individualism, computationalism, and folk psychology. Manuscript.

Challenges Burge's interpretations of the thought-experiments: e.g. twins have the same concept, neither of which is the public concept of arthritis. With remarks on computationalism and Marr's theory.

Campbell, J. 1982. Extension and psychic state: Twin Earth revisited. *Philosophical Studies* 42:67-89.

Argues that natural kind terms are token-reflexive, with reference ultimately fixed to the underlying explanatory properties of the surface qualities of local matter.

Crane, T. 1991. All the difference in the world. *Philosophical Quarterly* 41:1-25.

Twins share the same concepts. Contra Putnam: essentialism is fallacious; contra Burge: speakers share beliefs, but one has false belief about meaning.

Cummins, R. 1991. Methodological reflections on belief. In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.

We shouldn't rely on intuitions about thought-experiments; we need an empirical theory about belief. Belief contents are distinct from sentence contents; we have to distinguish linguistic from psychological semantics.

Devitt, M. 1990. Meanings just ain't in the head. In (G. Boolos, ed) *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.

Against Searle's theory of internal intentionality. Searle's theory requires magic to grasp external contents internally.

Dretske, F. 1993. The nature of thought. *Philosophical Studies* 70:185-99.

Argues that thought is extrinsic, but it is not essentially social. A system without a linguistic community could have thoughts, if it had an appropriate learning history.

Elugardo, R. 1993. Burge on content. *Philosophy and Phenomenological Research* 53:367-84.

Contra Burge on sofas: oblique that-clauses can't identify the (wide) way that the subject thinks of sofas, which is idiosyncratic and inexpressible.

Forbes, G. 1987. A dichotomy sustained. *Philosophical Studies* 51:187-211.

Gives a Fregean account of belief semantics to handle the Burge cases, and argues that the *type* of a proposition may be internal even if the token itself is not. With remarks on the relevance to Grice's program.

Georgalis, N. 1999. Rethinking Burge's thought experiment. *Synthese* 118:145-64.

Horowitz, A. 1995. Putnam, Searle, and externalism. *Philosophical Studies* 81:27-69.

Argues for a moderate externalism by synthesizing Putnam and Searle: internal intension leaves extension indeterminate, but it specifies the facts relevant to filling in the indeterminacy.

Koethe, J. 1992. And they ain't outside the head either. *Synthese* 90:27-53.

Ludwig, K. 1993. Externalism, naturalism, and method. In (E. Villanueva, ed) *Naturalism and Normativity*. Ridgeview.

Ludwig, K. 1996. Duplicating thoughts. *Mind and Language* 11:92-102.

Mandelkar, S. 1991. An argument against the externalist account of psychological content. *Philosophical Psychology* 4:375-82.

Argues that conscious experience is required for intentional states, and that any external relations could be satisfied without this experience, so external relations cannot suffice for intentional content.

McCulloch, G. 1992. The spirit of twin earth. *Analysis* 52:168-174.

Various arguments against Crane 1991 on externalism.

McDowell, J. 1977. On the sense and reference of a proper name. *Mind*.

McGilvray, J. 1998. Meanings are syntactically individuated and found in the head. *Mind and Language*

13:225-280.

McKinsey, M. 1991. The internal basis of meaning. *Pacific Philosophical Quarterly* 72:143-69.

Argues that meaning is determined by a certain kind of internal state, involving de se cognitive attitudes. These states aren't shared by twins, but are still narrow in a strong sense.

McKinsey, M. 1993. Curing folk psychology of arthritis. *Philosophical Studies* 70:323-36.

McKinsey, M. 1994. Individuating beliefs. *Philosophical Perspectives* 8:303-30.

Owens, J. 1983. Functionalism and the propositional attitudes. *Nous* 17:529-49.

Functional organization doesn't determine attitude content, even if we include inputs and outputs.

Perry, J. 1979. The problem of the essential indexical. *Nous* 13:3-21.

Indexicals are essential to some beliefs, so belief cannot just be a relation to a proposition. Belief contents must be at least in part construed relative to a subject. Separate belief object and belief state.

Putnam, H. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7:131-193. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

What is in the head doesn't determine the reference of our thoughts: my twin on Twin Earth refers to XYZ where I refer to H₂O. Content is determined by environment and linguistic community as well as by internal stereotypes.

Putnam, H. 1987. Meaning, other people, and the world. In *Representation and Reality*. MIT Press.

Meanings *still* aren't in the head.

Searle, J. R. 1983. *Intentionality*. Cambridge University Press.

Sure, meanings *are* in the head -- e.g. the content of a given visual experience is "the thing that is causing this experience".

Sosa, E. 1991. Between internalism and externalism. In (E. Villanueva, ed) *Consciousness*. Ridgeview.

Sosa, E. 1993. Abilities, concepts, and externalism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

On concepts as abilities, and on construals of abilities that lead to internalism and externalism. Maybe the relevant abilities are characterized externally but determined internally. Remarks on Putnam, Davidson,

Burge.

Stalnaker, R. 1993. Twin earth revisited. *Proceedings of the Aristotelian Society* 63:297-311.

Making sense of twin earth intuitions with an information-theoretic account of content: information depends on relations in normal conditions, which are extrinsic. With remarks on the context-sensitivity of content-attribution.

Wikforss, A. 2001. Social externalism and conceptual errors. *Philosophical Quarterly* 203:217-31.

Woodfield, A. 1982. Thought and the social community. *Inquiry* 25:435-50.

Burge's arguments show only that context-ascription is pragmatically sensitive to context, depending on the epistemic predicament of the ascriber. Content itself is still internal.

Zemach, E. M. 1976. Putnam's theory on the reference of substance terms. *Journal of Philosophy* 73:116-27.

Argues that the extension of 'water' is the same on earth and twin earth, using arguments from isotopes and scientific development. Molar properties determine classification. Remarks on historicism and the division of labor.

2.2b Externalism and Psychological Explanation (Burge, Fodor)

Adams, F. & K. Aizawa, 2001. The bounds of cognition. *Philosophical Psychology* 14:43-64.

Arjo, D. 1996. Sticking up for Oedipus: Fodor on intentional generalizations and broad content. *Mind and Language* 11:231-45.

Buller, D. J. 1992. "Narrow"-minded breeds inaction. *Behavior and Philosophy* 20:59-70.

Buller, D. J. 1997. Individualism and evolutionary psychology (or: In defense of "narrow" functions). *Philosophy of Science* 64:74-95.

Burge, T. 1982. Two thought experiments reviewed. *Notre Dame Journal of Formal Logic* 23:284-94.

Reply to Fodor 1982, clarification of position.

Burge, T. 1986. Individualism and psychology. *Philosophical Review* 95:3-45.

Psychology should be and is done non-individualistically, i.e. with reference to environment. Examples from vision, e.g. Marr.

Clark, A. & Chalmers, D. J. 1998. The extended mind. *Analysis* 58:7-19.

Advocates a different sort of "active externalism", based on the role of the environment in actively driving cognition. Beliefs can extend into an agent's immediate environment (e.g. a notebook) in this way.

Dretske, F. 1992. What isn't wrong with folk psychology. *Metaphilosophy* 23:1-13.

Argues that extrinsic properties can play a respectable role in scientific explanation; e.g. the histories of plants, animals, and devices are relevant in explaining their current behavior.

Egan, F. 1991. Must psychology be individualistic? *Philosophical Review* 100:179-203.

Maybe, maybe not. Contra Fodor: science can be non-individualistic. Contra Burge re oblique ascriptions and Marr.

Fodor, J. A. 1980. Methodological solipsism as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3:63-109. Reprinted in *RePresentations* (MIT Press, 1980).

Should do psychology without reference to the external world. What counts for psychology is in the head; who cares about truth, reference, and the rest?

Fodor, J. A. 1982. Cognitive science and the twin-earth problem. *Notre Dame Journal of Formal Logic* 23:98-118.

Twin Earth isn't a problem for cognitive science. Intents of utterances, de re/de dicto, etc. Truth conditions aren't in the head, but that's no problem.

Gauker, C. 1987. Mind and chance. *Canadian Journal of Philosophy* 17:533-52.

Globus, G. 1984. Can methodological solipsism be confined to psychology? *Cognition and Brain Theory* 7:233-46.

Methodological solipsism implies epistemological solipsism.

Hardcastle, V. G. 1997. [Explanation] is explanation better. *Philosophy of Science* 64:154-60.

Hurley, S. L. 1998. Vehicles, contents, conceptual structure, and externalism. *Analysis* 58:1-6.

Jacob, P. 1993. Externalism and the explanatory relevance of broad content. *Mind and Language* 8:131.

Kitcher, P. S. 1984. Narrow taxonomy and wide functionalism. *Philosophy of Science* 52:78-97.

Argues against Stich, Fodor, Block: use different taxonomies (narrow/wide) for different purposes. Both

are OK, functionalism *and* content survive.

Kobes, B. 1989. Semantics and psychological prototypes. *Pacific Philosophical Quarterly* 70:1-18.

Relates the individualism debate to Roschian prototype research.

Losonsky, M. 1995. Emdded systems vs. individualism. *Minds and Machines* 5:357-71.

Macdonald, C. 1992. Weak externalism and psychological reduction. In (D Charles & K. Lennon, eds) *Reduction, Explanation and Realism*. Oxford University Press.

Marras, A. 1985. The Churchlands on methodological solipsism and computational psychology. *Philosophy of Science* 52:295-309.

MS doesn't rule out all use of content, just of wide content. Narrow content is OK. With remarks on folk psychology and computation.

Maloney, J. C. 1985. Methodological solipsism reconsidered as a research strategy in cognitive psychology. *Philosophy of Science* 52:451-69.

Various problems for computational psychology handling content. It shares the problems of a naturalistic psychology.

McClamrock, R. 1991. Methodological individualism considered as a constitutive principle of scientific inquiry. *Philosophical Psychology* 4:343-54.

McClamrock, R. 1995. *Existential Cognition: Computational Minds in the World*. University of Chicago Press.

Noonan, H. W. 1984. Methodological solipsism: A reply to Morris. *Philosophical Studies* 48:285-290.

Noonan, H. W. 1986. Russellian thoughts and methodological solipsism. In (J. Butterfield, ed) *Language, Mind, and Logic*. Cambridge University Press.

Noonan, H. W. 1990. Object-dependent thoughts and psychological redundancy. *Analysis* 51:1-9.

Noonan, H. W. 1993. Object-dependent thoughts: A case of superficial necessity but deep contingency? In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Object-dependent thoughts are redundant in psychological explanation, as an explanation applying to a hallucinator will work as well. But this needn't defeat externalism in general. With remarks on self-knowledge.

Patterson, S. 1990. The explanatory role of belief ascriptions. *Philosophical Studies* 59:313-32.

Uses examples to argue that in explaining behavior we often ascribe beliefs in an individualistic way, even in cases where individual and community use diverge. These contents are at least sometimes expressible.

Patterson, S. 1991. Individualism and semantic development. *Philosophy of Science* 58:15-35.

Developmental psychologists attribute concepts individualistically.

Peacocke, C. 1993. Externalist explanation. *Proceedings of the Aristotelian Society* 67:203-30.

Externalist states are required for the explanation of relational properties. Counters objections from conceptual connections and dormitive-virtue worries, and applies to teleology, self-knowledge, etc.

Petrie, B. 1990. Nonautonomous psychology. *Southern Journal of Philosophy* 28:539-59.

Argues that behavior is often individuated widely for explanatory purposes, so that wide content is relevant, and that there is more to causation than local causation, so Stich's autonomy principle fails.

Pettit, P. 1986. Broad-minded explanation and psychology. In (P. Pettit & J. McDowell, eds) *Subject, Thought and Context*. Oxford University Press.

Rowlands, M. 1995. Against methodological solipsism: The ecological Approach. *Philosophical Psychology* 8:5-24.

Segal, G. 1989. The return of the individual. *Mind* 98:39-57.

Sterelny, K. 1990. Animals and individualism. In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

Stich, S. P. 1978. Autonomous psychology and the belief/desire thesis. *Monist* 61:573-91. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Beliefs are not in the head, so aren't good for psychological explanation. Interesting, but confuses the role of truth-values with truth-conditions.

Tuomela, R. 1989. Methodological solipsism and explanation in psychology. *Philosophy of Science* 56:23-47.

Wallace, J. & Mason, H. E. 1990. On some thought experiments about mind and meaning. In (C. Anderson & J. Owens, eds) *Propositional Attitudes*. CSLI.

Wilson, R. A. 1994. Causal depth, theoretical appropriateness, and individualism in psychology.

Philosophy of Science 61:55-75.

Wilson, R. A. 1995. *Cartesian Psychology and Physical Minds: Individualism and the Sciences of the Mind*. Cambridge University Press.

2.2c Externalism and Mental Causation

Adams, F. 1993. Fodor's modal argument. *Philosophical Psychology* 6:41-56.

Allen, C. 1995. It isn't what you think: A new idea about intentional causation. *Nous* 29:115-26.

Baker, L. R. 1994. Content and context. *Philosophical Perspectives* 8:17-32.

Argues contra Fodor that broad contents can be explanatory -- if they can't, no relational properties can.
Fodor's "no-conceptual-connection" and "cross-context" tests for causal powers fail to do the job.

Barrett, J. 1997. Individualism and the cross-contexts test. *Pacific Philosophical Quarterly* 78:242-??.

Braun, D. 1991. Content, causation, and cognitive science. *Australasian Journal of Philosophy* 69:375-89.

Arguments for the causal significance of broad content. Physical twins can differ in causal powers; broad content figures in (ceteris paribus) causal generalizations; can make twin arguments against narrow content too. Hmm.

Burge, T. 1989. Individuation and causation in psychology. *Pacific Philosophical Quarterly* 70:303-22.

Contra Fodor: psychological processes can play differing causal roles, despite being physically identical.

Burge, T. 1995. Intentional properties and causation. In (C. Macdonald & G. Macdonald, eds) *Philosophy of Psychology: Debates about Psychological Explanation*. Blackwell.

Reply to Fodor 1991.

Butler, K. 1996. Content, causal powers, and context. *Philosophy of Science* 63:105-14.

Christensen, D. 1992. Causal powers and conceptual connections. *Analysis* 52:163-8.

Fodor's modal argument for narrow content rests on a false analogy between cases concerning thoughts and those concerning planets.

Fodor, J. A. 1991. A modal argument for narrow content. *Journal of Philosophy* 88:5-26.

On when a difference in effects amounts to a difference in causal powers: when the effects are connected contingently, not conceptually, to the causes. Differences in wide content don't satisfy this, so aren't causal powers.

Garcia-Carpintero, M. 1994. The supervenience of mental content. *Proceedings of the Aristotelian Society* 68:117-135.

Mental content can be extrinsic and efficacious. Narrow content strategies don't work, as observation concepts are still extrinsic. One can't screen off the intrinsic part from the rest. Thought-experiments are inconclusive.

Heil, J. & Mele, A. 1991. Mental causes. *American Philosophical Quarterly* 28:61-71.

Reconciling Twin Earth with the causal relevance of content. Historical factors can be causally relevant.

Jacob, P. 1992. Externalism and mental causation. *Proceedings of the Aristotelian Society* 66:203-19.

Argues that externalist content is not causally efficacious, but is relevant to causal explanations of behavior indirectly, via the cognitive activities of others external to the system.

Klein, M. 1996. Externalism, content, and causation. *Proceedings of the Aristotelian Society* 96:159-76.

Lalor, B. J. 1997. It is what you think: intentional potency and anti-individualism. *Philosophical Psychology* 10:165-78.

Ludwig, K. 1993. Causal relevance and thought content. *Philosophical Quarterly* 44:334-53.

McGinn, C. 1991. Conceptual causation. *Mind* 100:525-46.

Montgomery, R. 1995. Non-Cartesian explanations meet the problem of mental causation. *Southern Journal of Philosophy* 33:221-41.

Owens, J. 1993. Content, causation, and psychophysical supervenience. *Philosophy of Science* 60:242-61.

Russow, L. M. 1993. Fodor, Adams, and causal properties. *Philosophical Psychology* 6:57-61.

Saidel, E. 1994. Content and causal powers. *Philosophy of Science* 61:658-65.

Segal, G. & Sober, E. 1991. The causal efficacy of content. *Philosophical Studies* 63:1-30.

Seymour, D. 1993. Some of the difference in the world: Crane on intentional causation. *Philosophical Quarterly* 43:83-89.

Sturgeon, S. 1994. Good reasoning and cognitive architecture. *Mind and Language* 9:88-101.

Epistemology requires the causal relevance of content, and the relevant content is narrow. On how various architectures might support this causal relevance, by being realized by more specific intrinsic features.

van Gulick, R. 1989. Metaphysical arguments for internalism and why they don't work. In (S. Silvers, ed) *ReRepresentation*. Kluwer.

Against some arguments for internalism: local causation doesn't imply local type-individuation, as distal relations affect distal causes and effects; and processes can have access to semantic properties via formal properties.

Wilson, R. A. 1992. Individualism, causal powers, and explanation. *Philosophical Studies* 68:103-39.

Science frequently appeals to relational and historical taxonomies, so either causal powers can be non-intrinsic or science needn't taxonomize by causal powers. With remarks on causal properties and conceptual connections.

Wilson, R. A. 1993. Against a priori arguments for individualism. *Pacific Philosophical Quarterly* 74:60-79.

Arguments from causal powers beg the question, either on whether relational properties can have causal powers or on whether science taxonomizes by causal powers, as relational properties are common in scientific explanation.

Yablo, S. 1997. Wide causation. *Philosophical Perspectives* 11:251-81.

2.2d

Externalism and the Theory of Vision

Burge, T. 1986. Individualism and psychology. *Philosophical Review* 95:3-45.

Psychology should be and is done non-individualistically, i.e. with reference to environment. Examples from vision, e.g. Marr.

Butler, K. 1996. Individualism and Marr's computational theory of vision. *Mind and Language* 11:313-37.

Butler, K. 1996. Content, computation, and individualism in vision theory. *Analysis* 56:146-54.

Cain, M. J. 2000. Individualism, twin scenarios and visual content. *Philosophical Psychology* 13:441-463.

- Davies, M. 1991. Individualism and perceptual content. *Mind* 100:461-84.
- Egan, F. 1992. Individualism, computation, and perceptual content. *Mind* 101:443-59.
- Egan, F. 1996. Intentionality and the theory of vision. In (K. Akins, ed) *Perception*. Oxford University Press.
- Francescotti, R. M. 1991. Externalism and the Marr theory of vision. *British Journal for the Philosophy of Science* 42:227-38.
- Kitcher, P. S. 1988. Marr's computational theory of vision. *Philosophy of Science* 55:1-24.
- Morton, P. 1993. Supervenience and computational explanation in vision theory. *Philosophy of Science* 60:86-99.
- Patterson, S. 1996. Success-orientation and individualism in the theory of vision. In (K. Akins, ed) *Perception*. Oxford University Press.
- Segal, G. 1989. Seeing what is not there. *Philosophical Review* 97:189-214.

Contra Burge, Marr's theory is individualistic. Intentional contents therein are neutral between twins' environments; nothing grounds a more specific attribution.

- Segal, G. 1991. Defence of a reasonable individualism. *Mind* 100:485-94.
- Shapiro, L. A. 1993. Content, kinds, and individualism in Marr's theory of vision. *Philosophical Review* 102:489-513.

Contra Segal, Marr's theory is non-individualistic even though it may classify twins together. Computational-level task descriptions rather than behavior guide content ascription, so the environment plays a crucial role.

- Shapiro, L. A. 1997. A clearer vision. *Philosophy of Science* 64:131-53.
- Shapiro, L. A. 1997. Junk representations. *British Journal for the Philosophy of Science* .

2.2e

Externalism and Computation

- Andler, D. 1995. Can we knock off the shackles of syntax? In (E. Villanueva, ed) *Content*. Ridgeview.

Butler, K. 1998. Content, computation, and individuation. *Synthese* 114:277-92.

Egan, F. 1995. Computation and content. *Philosophical Review* 104:181-203.

Egan, F. 1999. In defence of narrow mindedness. *Mind and Language* 14:177-94.

Kazez, J. R. 1994. Computationalism and the causal role of content. *Philosophical Studies* 75:231-60.

Kobes, B. 1990. Individualism and artificial intelligence. *Philosophical Perspectives* 4:429-56.

Winograd's SHRDLU doesn't support individualism: its concepts are anchored (to a fictional world) via its programmer, and it could have made errors.

Miscevic, N. 1996. Computation, content, and cause. *Philosophical Studies* 82:241-63.

Peacocke, C. 1995. Content, computation, and externalism. In (E. Villanueva, ed) *Content*. Ridgeview.

Peacocke, C. 1999. Computation as involving content: A response to Egan. *Mind and Language* 14:195-202.

Seager, W. E. 1992. Thought and syntax. *Philosophy of Science Association* 1992, 1:481-91.

Syntax is extrinsically determined, as well as semantics. So if broad content is irrelevant to psychology, syntax is too.

Wilson, R. A. 1994. Wide computationalism. *Mind* 103:351-72.

2.2f

Externalism and Self-Knowledge

Berg, J. 1998. First-person authority, externalism, and wh-knowledge. *Dialectica* 52:41-44.

Bernecker, S. 1996. Davidson on first-person authority and externalism. *Inquiry* 39:121-39.

Bernecker, S. 1996. Externalism and the attitudinal component of self-knowledge. *Nous* 30:262-75.

Bernecker, S. 1998. Self-knowledge and closure. In (P. Ludlow & N. Martin, eds) *Externalism and Self-Knowledge*. CSLI.

Bilgrami, A. 1992. Can externalism be reconciled with self-knowledge? *Philosophical Topics* 20:233-68.

Boghossian, P. 1989. Content and self-knowledge. *Philosophical Topics* 17:5-26.

We can't know our thought-contents by inference (circular), by introspection (because they're relational), or directly, so we can't know them at all.

Boghossian, P, 1992. Externalism and inference. *Philosophical Issues* 2:11-28.

Boghossian, P. 1994. The transparency of mental content. *Philosophical Perspectives* 8:33-50.

Boghossian, P. 1997. What the externalist can know a priori. *Proceedings of the Aristotelian Society* 97:161-75.

Brown, J. 1995. The incompatibility of anti-individualism and privileged access. *Analysis* 55:149-56.

Brown, J. 2000. Critical reasoning, understanding and self-knowledge. *Philosophy and Phenomenological Research* 61:659-676.

Brueckner, A. 1990. Scepticism about knowledge of content. *Mind* 99:447-51.

Brueckner, A. 1992. What an anti-individualist knows a priori. *Analysis* 52:111-18.

Contra McKinsey 1991, anti-individualism doesn't lead to a priori knowledge. The belief that water is wet doesn't conceptually entail facts about the external world (e.g. H₂O), although it may metaphysically necessitate them.

Brueckner, A. 1992. Semantic answers to skepticism. *Pacific Philosophical Quarterly* 73:200-19.

Brueckner, A. 1993. Skepticism and externalism. *Philosophia* 22:169-71.

Brueckner, A. 1994. Knowledge of content and knowledge of the world. *Philosophical Review*:103-327-43.

Brueckner, A. 1995. Trying to get outside your own skin. *Philosophical Topics* 23:79-111.

Brueckner, A. 1997. Externalism and memory. *Pacific Philosophical Quarterly* 78:1-12.

Brueckner, A. 1997. Is scepticism about self-knowledge incoherent? *Analysis* 4:287-90.

Brueckner, A. 2000. Externalism and the a prioricity of self-knowledge. *Analysis* 60:132-136.

Burge, T. 1988. Individualism and self-knowledge. *Journal of Philosophy* 85:649-63.

Knowledge of our thoughts is compatible with externalism: its content is self-referential and self-verifying. We needn't be able to explicate the content or its enabling conditions, or rule out twin possibilities.

Burge, T. 1996. Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society* 96:91-116.

Burge, T. 1998. Memory and self-knowledge. In (P. Ludlow & N. Martin, eds) *Externalism and Self-Knowledge*. CSLI.

Butler, K. 1997. Externalism, internalism, and knowledge of content. *philosophy and Phenomenological Research* 57:773-800.

Butler, K. 1998. Externalism and skepticism. *Dialogue* 37:13-34.

Chase, J. 2001. Is externalism about content inconsistent with internalism about justification? *Australasian Journal of Philosophy* 79:227-46.

Davidson, D. 1987. Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*.

Davies, M. 1998. Externalism, architecturalism, and epistemic warrant. In (C. Wright, B. Smith, and C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Ebbs, G. 1996. Can we take our words at face value? *Philosophy and Phenomenological Research* 56:499-530.

Edwards, J. 1998. The simple theory of colour and the transparency of sense experience. In (C. Wright, B. Smith, and C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Falvey, K. & Owens, J. 1994. Externalism, self-knowledge, and skepticism. *Philosophical Review* 103:107-37.

Falvey, K. 2000. The compatibility of anti-individualism and privileged access. *Analysis* 60:137-142.

Gallois, A. 1994. Deflationary self-knowledge. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind: The Place of Philosophy in the Study of Mind*. Kluwer.

Gallois, A. & O'Leary-Hawthorne, J. 1996. Externalism and skepticism. *Philosophical Studies* 81:1-26.

Externalist anti-skeptical arguments fail as they require us to know a priori that our terms designate natural kinds, and also because they require us to know a priori that externalism is true. A thorough analysis.

Georgalis, N. 1990. No access for the externalist: Discussion of Heil's "Privileged access". *Mind* 100:101-8.

Georgalis, N. 1994. Asymmetry of access to intentional states. *Erkenntnis* 40:185-211.

Gibbons, J. 1996. Externalism and knowledge of content. *Philosophical Review* 105:287-310.

Gibbons, J. 2001. Externalism and knowledge of the attitudes. *Philosophical Quarterly* 51:13-28.

Glock, H. J. & Preston, J. M. 1995. Externalism and first-person authority. *Monist* 78:515-33.

Goldberg, S. 1997. Self-ascription, self-knowledge, and the memory argument. *Analysis* 57:211-19.

Goldberg, S. 1999. The relevance of discriminatory knowledge of content. *Pacific Philosophical Quarterly* 80:136-56.

Goldberg, S. 1999. The psychology and epistemology of self-knowledge. *Synthese* 118:165-201.

Goldberg, S. 2000. Externalism and authoritative knowledge of content: A new incompatibilist strategy. *Philosophical Studies* 100:51-79.

Hall, L. 1998. The self-knowledge that externalists leave out. *Southwest Philosophy Review* 14.

Heal, J. 1998. Externalism and memory. *Aristotelian Society Supplement* 72:77-94.

Heil, J. 1988. Privileged access. *Mind* 98:238-51.

Kobes, B. 1996. Mental content and hot self-knowledge. *Philosophical Topics* 24:71-99.

LePore, E. 1990. Subjectivism and environmentalism. *Inquiry* 33:197-214.

Subjectivism and environmentalism seem to clash on knowledge of content, but it's OK: under environmentalism we still know our contents w/o evidence.

Ludlow, P. 1995. Externalism, self-knowledge, and the prevalence of slow-switching. *Analysis* 55:45-49.

Argues that cases of switching between language communities are quite common, so that Warfield's case for externalist self-knowledge doesn't work.

Ludlow, P. 1995. Social externalism, self-knowledge, and memory. *Analysis* 55:157-59.

Ludlow, P. 1995. Social externalism and memory: A problem? *Acta Analytica* 10:69-76.

Ludlow, P. 1997. On the relevance of slow switching. *Analysis* 57:285-86.

Ludlow, P. & Martin, N. 1998. *Externalism and Self-Knowledge*. CSLI.

Macdonald, C. 1995. Externalism and first-person authority. *Synthese* 104:99-122.

On reconciling externalism with the non-evidential character of first-person knowledge.

Macdonald, C. 1998. Externalism and authoritative self-knowledge. In (C. Wright, P. Smith, & C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Macdonald, C. , Smith, P. & Wright, C. 1998. *Knowing Our Own Minds: Essays in Self-Knowledge*. Oxford University Press.

McKinsey, M. 1987. Apriorism in the philosophy of language. *Philosophical Studies* 52:1-32.

Argues that we can know the meaning of our words a priori. Analyzes twin earth cases by separating propositional meaning from linguistic meaning, which is indexical, fixes reference, and is knowable a priori.

McKinsey, M. 1991. Anti-individualism and privileged access. *Analysis* 51:9-16.

Contra Burge: if there are conceptual connections between wide contents and and the external world, then we can't know wide contents a priori, as otherwise we could know a priori that the world exists.

McKinsey, M. 1994. Accepting the consequences of anti-individualism. *Analysis* 54:124-8.

Reply to Brueckner 1992: The claim that belief metaphysically necessitate external facts is trivial. Almost all states do that, for Kripkean reason.

McLaughlin, B. P. & Tye, M. 1998. Externalism, Twin Earth, and self-knowledge. In (C. Macdonald, P. Smith, & C. Wright, eds) *Knowing Our Own Minds: Essays in Self-Knowledge*. Oxford University Press.

McLaughlin, B. P. , & Tye, M. 1998. Is content-externalism compatible with privileged access? *Philosophical Review* 107:349-380.

Miller, R. W. 1997. Externalist self-knowledge and the scope of the a priori. *Analysis* 57:67-74.

Peacocke, C. 1996. Entitlement, self-knowledge, and conceptual redeployment. *Proceedings of the Aristotelian Society* 96:117-58.

Raffman, D. 1998. First-person authority and the internal reality of beliefs. In (C. Wright, B. Smith, & C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Sawyer, S. 1998. Privileged access to the world. *Australasian Journal of Philosophy* 76:523-533.

Schiffer, S. 1992. Boghossian on externalism and inference. *Philosophical Issues* 2:29-38.

Szubka, T. 2000. Meaning rationalism, a priori, and transparency of content. *Philosophical Psychology* 13:491-503.

Tye, M. 1998. Externalism and memory. *Aristotelian Society Supplement* 72:77-94.

Warfield, T. A. 1992. Privileged self-knowledge and externalism are compatible. *Analysis* 52:232-37.

Boghossian's argument that externalism threatens self-knowledge fails: twin cases needn't be relevant alternatives (unless they are actual), so they don't threaten knowledge of content, by the usual standards of knowledge.

Warfield, T. A. 1995. Knowing the world and knowing our minds. *Philosophy and Phenomenological Research*.

Argues that externalism and self-knowledge imply the falsity of skepticism (though externalism alone does not). And arguments against externalist self-knowledge are no better than standard skeptical arguments.

Warfield, T. A. 1997. Externalism, privileged self-knowledge, and the irrelevance of slow switching. *Analysis* 57:282-84.

Wyler, T. 1994. First-person authority and singular thoughts. *Zeitschrift fur Philosophie Forschung* 48:585-94.

2.2g

The Status of Narrow Content

Adams, F. , Drebusenko, D. , Fuller, G. & Stecker, R. 1990. Narrow content: Fodor's folly. *Mind and Language* 5:213-29.

Traces and criticizes Fodor's position on narrow content. Argues that narrow content isn't content, and doesn't explain behavior. Fun but arguable.

Adams, F. & Fuller, G. 1992. Names, contents, and causes. *Mind and Language* 7:205-21.

Argues that problems with names don't require an appeal to narrow content in explanation. Broad content plus associated descriptions will do the job.

Antony, L. 1989. Semantic anorexia: On the notion of content in cognitive science. In (G. Boolos, ed) *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.

Representational cognitive science has no need for narrow content -- wide contents and formal properties can do all the work. Argues that the semantics of mental expressions needn't mirror the semantics of language.

Aydede, M. 1997. Has Fodor really changed his mind on narrow content? *Mind and Language* 12:422-58.

Baker, L. R. 1985. A farewell to functionalism. *Philosophical Studies* 48:1-14.

Argues that type-identical functional states can differ in narrow content, so methodological solipsism fails. Uses the example of identical programs for playing chess and arms negotiations.

Baker, L. R. 1985. Just what do we have in mind? *Midwest Studies in Philosophy* 10:25-48.

Some implausible twin cases trying to show that mental life can vary wildly while preserving physical/computational state. Bizarre.

Baker, L. R. 1986. Content by courtesy. *Journal of Philosophy* 84:197-213.

Baker, L. R. 1987. *Saving Belief*. Princeton University Press.

Lots of arguments against narrow content. Very stimulating, though wrong.

Biro, J. I. 1992. In defense of social content. *Philosophical Studies* 67:277-93.

Contra Loar 1988, the contents of "that"-clauses often reflects psychological content, even if it sometimes does not. We don't need narrow content.

Block, N. 1991. What narrow content is not. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

There are big problems specifying the "mapping" and the relevant contexts for Fodor's theory noncircularly. Narrow content either collapses into syntax or is too coarse-grained. Nontrivial narrow content must be holistic.

Block, N. 1995. Ruritania revisited. In (E. Villanueva, ed) *Content*. Ridgeview.

Brown, C. 1993. Belief states and narrow content. *Mind and Language* 8:343-67.

Criticizes the "bracketing" strategy of Stich and Walker, and argues that intrinsic belief state should be individuated according to how it embeds in different environments. With a comparison with Fodor's related theory.

Chalmers, D. J. 1994. The components of content. Manuscript.

Argues for a two-dimensional intensional theory, with different kinds of intensions constituting notional and relational content. Notional content governs the dynamics of thought and behavior, and is primary in explanation.

Davies, M. 1986. Externality, psychological explanation, and narrow content. *Proceedings of the Aristotelian Society* 60:263-83.

Comments on Fodor 1987. Fodor doesn't make a conclusive case against externalism; but narrow content may be promising, and inexpressibility doesn't pose any real problems. With comparisons to neo-Fregean theories.

Dennett, D. C. 1983. Beyond belief. In (A. Woodfield, ed) *Thought and Object*. Oxford University Press. Reprinted in *The Intentional Stance* (MIT Press, 1987).

What matters are not propositional attitudes but notional attitudes; but it's hard to calibrate notional worlds. Very nice.

Devitt, M. 1990. The narrow representational theory of mind. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Not syntactic psychology nor wide psychology, but narrow psychology.

Field, H. 1989. "Narrow" aspects of intentionality and the information-theoretic approach to content. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Fodor, J. A. 1987. Individualism and supervenience. In *Psychosemantics*. MIT Press.

Science taxonomizes by causal powers, which are locally supervenient, so psychology needs a narrow notion of content. Proposes that a relativized notion -- a function from context to extension -- can do the job. Nice.

Jackson, F. , and Pettit, P. 1993. Some content is narrow. In (J. Heil and A. Mele, eds) *Mental Causation*. Oxford University Press.

Argues that folk psychology needs a notion of narrow content to provide robust predictive behavioral generalizations that covers doppelgangers. If not, then some behavioral patterns would be fluke.

LePore, E. & Loewer, B. 1986. Solipsistic semantics. *Midwest Studies in Philosophy* 10:595-614.

There's no good way to construe narrow content. Phenomenologist strategy is intrinsically wide, and indexicalist strategy can't specify content.

LePore, E. & Loewer, B. 1989. Dual aspect semantics. In (S. Silvers, ed) *ReRepresentation*. Kluwer.

Loar, B. 1987. Social content and psychological content. In (R. Grimm & D. Merrill, eds) *Contents of Thought*. University of Arizona Press.

Uses examples to argue that psychological content is not fixed by the content of "that"-clauses in belief ascription, and vice versa. We require a subtler kind of narrow content to capture what's going on.

Loar, B. 1987. Subjective intentionality. *Philosophical Topics* 15:89-124.

Maloney, J. C. 1991. Saving psychological solipsism. *Philosophical Studies* 61:267-83.

Contests the "provoked/aggravated assault" example of Baker 1986. If they're doppelgangers, then their narrow content can't differ.

Manfredi, P. 1993. Two routes to narrow content: both dead ends. *Philosophical Psychology* 6:3-22.

McDermott, M. 1986. Narrow content. *Australasian Journal of Philosophy* 64:277-88.

Narrow beliefs are de re beliefs about our inputs and outputs.

McGilvray, J. 1998. Meanings are syntactically individuated and found in the head. *Mind and Language* 13:225-280.

Putnam, H. 1987. Fodor and Block on narrow content. In *Representation and Reality*. MIT Press.

Against perceptual-prototype and conceptual-role accounts of narrow content.

Quillen, K. 1986. Propositional attitudes and psychological explanation. *Mind and Language* 1:133-57.

Can't get a 'mode of presentation' account of narrow content to work, either through description theory or prototypes. Psych should be non-individualist.

Recanati, F. 1990. Externalism and narrow content. *Nous*.

There are levels of narrowness, varying by whether independence is of actual or normal environment. Argues that this can be consistent with externalism.

Recanati, F. 1994. How narrow is narrow content? *Dialectica* 48:209-29.

Schiffer, S. 1989. Fodor's character. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Segal, G. 2000. *A Slim Book about Narrow Content*. MIT Press.

Silverberg, A. 1995. Narrow content: A defense. *Southern Journal of Philosophy* 33:109-27.

Stalnaker, R. C. 1990. Narrow content. In (C. A. Anderson & J. Owens, eds) *Propositional Attitudes*. CSLI.

On some problems with narrow content, contra Loar 1987. Narrow content is hard to spell out with "diagonal" propositions. Loar doesn't show that psychological content is narrow. With some remarks on privileged access.

Stich, S. P. 1991. Narrow content meets fat syntax. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Argues that narrow content is still too coarse-grained for explanation, classifying psychologically distinct states together. Use syntax instead.

Taylor, K. 1989. Supervenience and levels of meaning. *Southern Journal of Philosophy* 27:443-58.

Argues that the partial character construal of narrow content is not interestingly semantic. It collapses into syntax or phenomenology.

Taylor, K. 1989. Narrow content functionalism and the mind-body problem. *Nous* 23:355-72.

Uses a "fraternal twin earth" thought experiment to show that even de dicto attributions don't supervene on narrow role, and that narrow content can't be explicated descriptively unless it collapses into phenomenalism.

Vaughan, R. 1989. Searle's narrow content. *Ratio* 2:185-90.

White, S. 1982. Partial character and the language of thought. *Pacific Philosophical Quarterly* 63:347-65.

Replies to Burge/Stich arguments by introducing partial character -- a function from context to content, analogous to Kaplan's character -- as the semantic property determined by functional state and relevant to explanation.

White, S. 1992. Narrow content and narrow interpretation. In *The Unity of the Self*. MIT Press.

Argues for an account of narrow content in terms of notional worlds, by considering "objective optimality" across worlds. This allows for a sort of narrow radical interpretation. With arguments against Stalnaker.

Williams, M. 1990. Social norms and narrow content. *Midwest Studies in Philosophy* 15:425-462.

Narrow content theories can't handle the normativity of content. In-depth treatment of Burge cases and of the failures of causal and conceptual-role accounts. Normativity is fundamentally social. A long, interesting paper.

Williamson, T. 1998. The broadness of the mental: Some logical issues. *Philosophical Perspectives* 12:389-410.

2.2h

Miscellaneous

Brook, D. 1992. Substantial mind. *South African Journal of Philosophy* 1:15-21.

Brown, D. J. 1993. Swampman of La Mancha. *Canadian Journal of Philosophy* 23:327-48.

An entertaining fable about a swampthing doppelganger of a murder witness. Does he have content? With plot twists about personal identity.

Brown, D. J. 1996. A furry tile about mental representation. *Philosophical Quarterly* 185:448-66.

Buekens, F. 1994. Externalism, content, and causal histories. *Dialectica* 48:267-86.

de Vries, W. A. 1996. Experience and the swamp creature. *Philosophical Studies* 82:55-80.

Argues that a swampthing isn't intelligent or intentional, with different physiological processes and no sensations, as these are functional kinds.

Edwards, S. 1994. *Externalism in the Philosophy of Mind*. Avebury.

Engel, P. 1987. Functionalism, belief, and content. In (Torrance, ed) *The Mind and the Machine*. Horwood.

Gauker, C. 1991. Mental content and the division of epistemic labour. *Australasian Journal of Philosophy* 69:302-18.

Gibbons, J. 1993. Identity without supervenience. *Philosophical Studies* 70:59-79.

Houghton, D. 1997. Mental content and external representations: internalism, anti-internalism. *Philosophical Quarterly* 47:159-77.

Jackson, F. & Pettit, P. 1988. Functionalism and broad content. *Mind* 97:318-400.

Should construe functionalism broadly rather than narrowly; then can handle the problem of broad content.

Katz, J. 1990. The domino theory. *Philosophical Studies* 58:3-39.

Anti-intensional arguments are not independent but a series of dominos.

Quine/Quine/Davidson/Putnam/Burge rise and fall together.

Macdonald, C. 1990. Weak externalism and mind-body identity. *Mind* 99:387-404.

McCulloch, G. 1995. *The Mind and its World*. Routledge.

McGinn, C. 1982. The structure of content. In (A. Woodfield, ed) *Thought and Object*. Oxford University Press.

Belief content has two distinct elements, one causal-explanatory, the other truth-related.

Owens, J. 1987. In defense of a different Doppelganger. *Philosophical Review* 96:521-54.

Owens, J. 1992. Psychophysical supervenience: Its epistemological foundation. *Synthese* 90:89-117.

Pereboom, D. 1995. Conceptual structure and the individuation of content. *Philosophical Perspectives* 9:401-428.

Preti, C. 2000. Belief and desire under the elms. *Protosociology* 14:270-284.

Rey, G. 1992. Semantic externalism and conceptual competence. *Proceedings of the Aristotelian Society* 66:315-33.

Supplements externalist "locking" theories of content with an account of internal "conceptions" by which thoughts lock onto environmental kinds, with that aid of dthat operators, thus solving various philosophical problems.

Rowlands, M. 1995. Externalism and token-token identity. *Philosophia* 24:359-75.

Rowlands, M. 1999. *The Body in Mind: Understanding Cognitive Processes*. Cambridge University Press.

Rudd, A. 1997. Two types of externalism. *Philosophical Quarterly* 47:501-7.

Seager, W. E. 1992. Externalism and token identity. *Philosophical Quarterly* 42:439-48.

Stalnaker, R. C. 1989. On what's in the head. *Philosophical Perspectives* 3:287-319.

Thomas, J. 1996. Analogies and the mind of the replica: Sunburn, the little green bug, and the fake plant. *Philosophical Quarterly* 46:364-371.

Walker, V. 1990. In defense of a different taxonomy: A reply to Owens. *Philosophical Review* 99.

Contra Owens 1987: wide intentional descriptions and molar bodily descriptions don't exhaust the options. A bracketing strategy gives a narrow intentional taxonomy of mental states.

Williams, M. 1990. Externalism and the philosophy of mind. *Philosophical Quarterly* 40:352-80.

Woodfield, A. 1986. Two categories of content. *Mind and Language* 1:319-54.

2.3

Causal Theories of Content

2.3a

Information-Based Accounts (Dretske, etc)

Barwise, J. & Perry, J. 1983. *Situations and Attitudes*. MIT Press.

Barwise, J. 1986. Information and circumstance. *Notre Dame Journal of Formal Logic*.

Defending information against Fodor 1986. Information is objective but relational, and depends on the relevant constraints between representation and environment. Circumstances play a vital role.

Barwise, J. 1987. Unburdening the language of thought. *Mind and Language*.

Bogdan, R. J. 1988. Information and semantic cognition: An ontological account. *Mind and Language*.

From material (formal) info to semantic info via teleology; from semantic information to representation via internal structure. Cute. With a good reply by Israel, and a terse reply by Dretske.

Bogdan, R. J. 1987. Mind, content and information. *Synthese*.

Clark, A. 1993. Mice, shrews, and misrepresentation. *Journal of Philosophy* 90:290-310.

Uses information theory to analyze misrepresentation. A signal represents what it carries most information about, not what it correlates best with. Treating some signals as noise can increase information content.

Coulter, J. 1995. The informed neuron: Issues in the use of information theory in the behavioral sciences. *Minds and Machines* 5:583-96.

Dretske, F. 1981. *Knowledge and the Flow of Information*. MIT Press.

Defines knowledge content is in terms of information-flow from events, and applies to various aspects of psychology.

Dretske, F. 1983. *Precis of Knowledge and the Flow of Information*. *Behavioral and Brain Sciences* 6:55-90.

A summary of the book, with commentary and replies.

Dretske, F. 1990. Putting information to work. In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

On the causal role of information (as opposed to meaning). Information is causally efficacious if considered with respect to learning. With commentary by Brian Smith.

Fodor, J. A. 1986. Information and association. *Notre Dame Journal of Formal Logic* 27.

Transmission of information is no good without the encoding of information. With criticisms of associative networks, which transmit without encoding, and criticism of Barwise & Perry's account of attunement to a relation.

Fodor, J. A. 1987. A situated grandmother. *Mind and Language*.

Foley, R. 1987. Dretske's 'information-theoretic' account of knowledge. *Synthese*.

Gjelsvik, O. 1991. Dretske on knowledge and content. *Synthese* 86:425-41.

Grandy, R. 1987. Information-based epistemology, ecological epistemology and epistemology naturalized. *Synthese* 70:191-203.

Shannon's notion of information is more useful for naturalized epistemology than Dretske's.

Hardcastle, V. G. 1994. Indicator semantics and Dretske's function. *Philosophical Psychology* 7:367-82.

Heller, M. 1991. Indication and what might have been. *Analysis* 51:187-91.

We need to analyze indication in terms of "close enough" worlds; the relevant conditionals are "might"-conditionals.

Israel, D. & Perry, J. 1990. What is information? In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

Jackendoff, R. 1985. Information is in the mind of the beholder. *Linguistics and Philosophy* 8:23-33.

Argues that a representationalist theory of semantics beats a realist one.

Loewer, B. 1987. From information to intentionality. *Synthese*.

Morris, W. E. 1990. The regularity theory of information. *Synthese* 82:375-398.

Dretske has problems with ruling out alternative possibilities; and there is a gap between information-caused belief and knowledge.

Savitt, S. 1987. Absolute informational content. *Synthese* 70:185-90.

Makes a distinction between absolute information and information that's relative to other knowledge.

Sayre, K. M. 1986. Intentionality and information processing: An alternative model for cognitive science. *Behavioral and Brain Sciences* 9:121-38.

Sayre, K. M. 1987. Cognitive science and the problem of semantic content. *Synthese* 70:247-69.

On problems with a computational approach to content: computers process info(t), the non-semantic content of communication theory, not info(s), or semantic content. Get info(s) from efficient processing of mutual info(t).

Sturdee, D. 1997. The semantic shuffle: Shifting emphasis in Dretske's account of representational content. *Erkenntnis* 47:89-104.

Taylor, K. 1987. Belief, information and semantic content: A naturalist's lament. *Synthese* 71:97-124.

Usher, M. 2001. A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language* 16:331-334.

Winograd, T. 1987. Cognition, attunement and modularity. *Mind and Language*.

Zalabardo, J. L. 1995. A problem for information-theoretic semantics. *Synthese* 105:1-29.

2.3b Asymmetric Dependence (Fodor)

Fodor, J. A. 1987. Meaning and the world order. In *Psychosemantics*. MIT Press.

Defends and refines a causal theory, using the notion of asymmetric dependence of a token upon the world.

Fodor, J. A. 1990. A theory of content II. In *A Theory of Content*. MIT Press.

Defending the asymmetric dependence theory against various objections.

Adams, F. & Aizawa, K. 1992. 'X' means X: Semantics Fodor-style. *Minds and Machines* 2:175-83.

Adams, F. & Aizawa, K. 1993. Fodorian semantics, pathologies, and "Block's problem". *Minds and Machines* 3:97-104.

Adams, F. & Aizawa, K. 1994. 'X' means X: Fodor/Warfield semantics. *Minds and Machines* 4:215-31.

Adams, F. & Aizawa, K. 1997. Fodor's asymmetric causal dependency theory and proximal projections. *Southern Journal of Philosophy* 35:433-437.

Antony, L. & Levine, J. 1991. The nomic and the robust. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Baker, L. R. 1990. On a causal theory of content. *Philosophical Perspectives*.

Baker, L. R. 1991. Has content been naturalized? In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Bernier, P. 1993. Narrow content, context of thought, and asymmetric dependence. *Mind and Language* 8:327-42.

Boghossian, P. 1991. Naturalizing content. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Argues that Fodor's theory is a type-1 theory, requiring naturalistically specifiable circumstances in which a symbol is only caused by its referent; and that these theories fail for various reasons, e.g. verificationism.

Cram, H-R. 1992. Fodor's causal theory of representation. *Philosophical Quarterly* 42:56-70.

Fodor's theory has counterexamples and can't explain its counterfactuals; but we can save it by borrowing from Dretske's account of misrepresentation.

Gibson, M. 1996. Asymmetric dependencies, ideal conditions, and meaning. *Philosophical Psychology* 9:235-59.

Loar, B. 1991. Can we explain intentionality? In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Maloney, J. C. 1990. Mental misrepresentation. *Philosophy of Science* 57:445-58.

Manfredi, P. A. & Summerfield, D. M. 1992. Robustness without asymmetry: A flaw in Fodor's theory of content. *Philosophical Studies* 66:261-83.

Rupert, R. 2000. Dispositions indisposed: Semantic atomism and Fodor's theory of content. *Pacific Philosophical Quarterly* 81:325-349.

Seager, W. E. 1993. Fodor's theory of content: problems and objections. *Philosophy of Science* 60:262-77.

Wallis, C. 1995. Asymmetric dependence, representation, and cognitive science. *Southern Journal of Philosophy* 33:373-401.

Warfield, T. A. 1994. Fodorian semantics: A reply to Adams and Aizawa. *Minds and Machines* 4:205-14.

2.3c Causal Accounts, General

Aizawa, K. 1994. Lloyd's dialectical theory of representation. *Mind and Language* 9:1-24.

Cummins, R. 1989. Representation and covariation. In (S. Silvers, ed) *ReRepresentation*. Kluwer.

Cummins, R. 1997. The LOT of the causal theory of mental content. *Journal of Philosophy* 94:535-542.

Fodor, J. A. 1984. Semantics, Wisconsin style. *Synthese* 59:231-50. Reprinted in *RePresentations* (MIT Press, 1980).

A somewhat sympathetic commentary on the Dretske/Stampe causal theories, but raising the problem of misrepresentation.

Fodor, J. A. 1990. Information and representation. In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

Godfrey-Smith, P. 1989. Misinformation. *Canadian Journal of Philosophy* 19:533-50.

On various attempts to solve the error problem and why they don't work.

Godfrey-Smith, P. 1991. Signal, decision, action. *Journal of Philosophy* 88:709-22.

World-head reliability is just as important as head-world reliability. With arguments and examples from signal detection theory.

Jacquette, D. 1996. Lloyd on intrinsic natural representation in simple mechanical minds. *Minds and Machines* 6:47-60.

Maloney, J. C. 1994. Content: Covariation, control, and contingency. *Synthese* 100:241-90.

McLaughlin, B. P. 1987. What is wrong with correlational psychosemantics. *Synthese*.

Ray, G. 1997. Fodor and the inscrutability problem. *Mind and Language* 12:475-89.

Stampe, D. 1977. Towards a causal theory of linguistic representation. *Midwest Studies in Philosophy* 2:42-63.

Stampe, D. 1986. Verificationism and a causal account of meaning. *Synthese* 69:107-37.

Stampe, D. 1991. Content, context, and explanation. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Viger, C. D. 2001. Locking on to the language of thought. *Philosophical Psychology* 14:203-215.

Warmbrod, K. 1992. Primitive representation and misrepresentation. *Topoi* 11:89-101.

Weitzman, L. 1996. What makes a causal theory of content anti-skeptical? *Philosophy and Phenomenological Research* 56:299-318.

2.3d

Teleological Approaches (Millikan, etc)

Adams, F. & Aizawa, K. 1997. Rock beats scissors: Historicalism fights back. *Analysis* 57:273-81.

Agar, N. 1993. What do frogs really believe? *Australasian Journal of Philosophy* 71:1-12.

Argues that a teleological account can resolve content indeterminacies, by an appeal to counterfactuals in examining what properties were selected for.

Bogdan, R. 1994. *Grounds for Cognition: How Goal-Guided Behavior Shapes the Mind*. Lawrence

Erlbaum.

Braddon-Mitchell, D. & Jackson, F. 1997. The teleological theory of content. *Australasian Journal of Philosophy* 75:474-89.

Clarke, M. 1996. Darwinian algorithms and indexical representation. *Philosophy of Science* 63:27-48.

Dennett, D. C. 1988. Fear of Darwin's optimizing rationale. Manuscript.

Defends evolutionary theories of content against Fodor.

Dennett, D. C. 1988. Evolution, error and intentionality. In *The Intentional Stance*. MIT Press.

Attacks original intentionality (Fodor/Burge/Dretske/Searle/Kripke) -- our intentionality, if anything, is derived through evolution, and so is as indeterminate as that of an artifact.

Dretske, F. 1986. Misrepresentation. In (R. Bogdan, ed) *Belief: Form, Content, and Function*. Oxford University Press.

Tries to deal with misrepresentation by appealing to function.

Elder, C. L. 1998. What versus how in naturally selected representations. *Mind* 107:349-363.

Fodor, J. A. 1990. Psychosemantics, or, Where do truth conditions come from? In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Truth conditions are "entry conditions" for belief under "normal function". Later repudiated.

Fodor, J. A. 1990. A theory of content I. In *A Theory of Content*. MIT Press.

Teleological solutions can't work, because of underdetermination and so on.

Keeley, B. 1999. Fixing content and function in neurobiological systems: The neuroethology of electroreception. *Biology and Philosophy* 14:395-430.

Lalor, B. J. 1998. Swampman, etiology, and content. *Southern Journal of Philosophy* 36:215-232.

Levine, J. 1996. Swampjoe: mind or simulation? *Mind and Language* 11:86-91.

Macdonald, G. 1989. Biology and representation. *Mind and Language* 4:186-200.

Matthen, M. 1988. Biological functions and perceptual content. *Journal of Philosophy* 85:5-27.

Millikan, R. G. 1979. An evolutionist approach to language. *Philosophy Research Archives* 5.

Millikan, R. G. 1984. *Language, Thought and Other Biological Categories*. MIT Press.

An evolutionary account of thought, content, and various intentional phenomena, appealing to proper functions and adaptational role to individuate contents.

Millikan, R. G. 1986. Thoughts without laws: Cognitive science with content. *Philosophical Review* 95:47-80.

The content of a desire is its adaptational Proper Function; the content of a belief is its Normal Condition for success.

Millikan, R. G. 1989. Biosemantics. *Journal of Philosophy* 86:281-97.

Representation content is determined by the consumption of a representation, not its production. The representation-world correspondence is best taken as a normal condition for the consumer's function.

Millikan, R. G. 1989. In defense of proper functions. *Philosophy of Science* 56:288-302.

Millikan, R. G. 1990. Compare and contrast Dretske, Fodor, and Millikan on teleosemantics. *Philosophical Topics* 18:151-61.

Contrasting positions on the role of representation production and consumption; also on the role of reliability, articulateness, and learning.

Millikan, R. G. 1991. Speaking up for Darwin. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

A reply to some of Fodor's criticisms of teleological theories in _Psychosemantics_ and elsewhere. With some remarks on Fodor's asymmetric dependence theory.

Millikan, R. G. 1993. *White Queen Psychology and Other Essays for Alice*. MIT Press,

A collection of papers on teleological semantics and other issues about psychology and mental content.

Millikan, R. G. 1996. On swampkinds. *Mind and Language* 11:103-17.

Millikan, R. G. 1997. Troubles with Wagner's reading of Millikan. *Philosophical Studies* 86:93-96.

Neander, K. 1995. Misrepresenting and malfunctioning. *Philosophical Studies* 79:109-41.

Neander, K. 1995. Dretske's innate modesty. *Australasian Journal of Philosophy* 74:258-74.

Neander, K. 1996. Swampman meets swampcow. *Mind and Language* 11:118-29.

It's not unreasonable to deny a swampthing beliefs: swampcows aren't cows and swamphearts aren't hearts. Semantic norms are plausibly grounded in biological norms and so in history.

Newton, N. 1992. Dennett on intrinsic intentionality. *Analysis* 52:18-23.

Contra Dennett 1988, designed creatures can have intrinsic (if not original) intentionality. Overall purpose is dependent on designer's goals, but specific contents need not be.

Papineau, D. 1984. Representation and explanation. *Philosophy of Science* 51:550-72.

A teleological theory of belief/desire contents: the satisfaction conditions for a desire are those effects for which it was selected; truth conditions for a belief are circumstances resulting in satisfaction of desires.

Papineau, D. 1990. Truth and teleology. In (D. Knowles, ed) *Explanation and its Limits*. Cambridge University Press.

Best theory is combination of a success-guaranteeing account of truth-conditions with a teleological account of desire.

Papineau, D. 1991. Teleology and mental states. *Proceedings of the Aristotelian Society* 65.

Papineau, D. 1996. Doubtful intuitions. *Mind and Language* 11:130-32.

Papineau, D. 1998. Teleosemantics and indeterminacy. *Australasian Journal of Philosophy* 76:1-14.

Papineau, D. 2001. The status of teleosemantics, or how to stop worrying about Swampman. *Australasian Journal of Philosophy* 79:279-89.

Pickles, D. 1989. Intentionality, representation, and function. Sussex University, Cognitive Science Research Paper 140.

Combining the analysis-relative and historical accounts of function, and using these to give an account of intentionality: representation are produced by conditional productive functions. Argues against Fodor on indeterminacy.

Pietrowski, P. M. 1992. Intentionality and teleological error. *Pacific Philosophical Quarterly* 73:267-82.

Millikan's theory has an implausible consequence: creatures' belief contents can involve properties which they cannot discriminate. With examples.

Ross, D. & Zawidzki, T. 1994. Information and teleosemantics. *Southern Journal of Philosophy* 32:393-419.

Rountree, J. 1997. The plausibility of teleological content ascriptions: A reply to Pietroski. *Pacific Philosophical Quarterly* 78:404-??.

Rowlands, M. 1996. Teleological semantics. *Mind* 106:279-304.

Rupert, R. D. 1999. Mental representations and Millikan's theory of intentional content: Does biology chase causality? *Southern Journal of Philosophy* 37:113-140.

Sehon, S. R. 1994. Teleology and the nature of mental states. *American Philosophical Quarterly* 31:63-72.

Shapiro, L. 1996. Representation from bottom to top. *Canadian Journal of Philosophy* 26:523-42.

Shapiro, L. 1992. Darwin and disjunction: Foraging theory and univocal assignments of content. *Philosophy of Science Association* 1992, 1:469-80.

Sullivan, S. R. 1993. From natural function to indeterminate content. *Philosophical Studies* 69:129-37.

Wagner, S. 1996. Teleosemantics and the troubles of naturalism. *Philosophical Studies* 82:81-110.

Teleosemantics has big problems with indeterminacy, holism, false belief, and "psychophysical normalcy" in causation. So do all naturalistic stories.

Zawidzki, T. & Ross, D. 1994. Information and teleosemantics. *Southern Journal of Philosophy* 32:393-419.

2.3e Conceptual Role Approaches

Block, N. 1986. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* 10:615-78.

An in-depth program for conceptual-role semantics, and its role in a two-factor account of meaning. Also a defense of narrow content.

Block, N. 1988. Functional role and truth conditions. *Proceedings of the Aristotelian Society* 61:157-181.

A defense of functional role semantics, and an account of its relation to truth-conditional factors. A two-factor theory will handle wide content.

Boghossian, P. A. 1994. Inferential-role semantics and the analytic/synthetic distinction. *Philosophical Studies*.

No matter how we understand the denial of the analytic/synthetic distinction, the falsity of inferential-role semantics does not follow. The meaning-constitutive inferences needn't be the analytic inferences.

Brandom, R. 1994. *Making It Explicit*. Harvard University Press.

Brandom, R. 1994. Reasoning and representing. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind*. Kluwer.

Cummins, R. 1992. Conceptual role semantics and the explanatory role of content. *Philosophical Studies* 65:103-127.

CRS conflates representation content and attitude content (which depends on a representation's "target"), so can't handle representation content; it makes all content-based explanations vacuous; and it can't handle error properly.

Field, H. 1977. Logic, meaning, and conceptual role. *Journal of Philosophy* 74:379-409.

Explicates conceptual role in terms of conditional probability, and analyzes meaning as conceptual role plus reference. With remarks on truth, descriptions, and synonymy.

Field, H. 1978. Mental representation. *Erkenntnis* 13:9-61.

Fodor, J. A. & LePore, E. 1991. Why meaning (probably) isn't conceptual role. *Mind and Language* 6:328-43.

Conceptual role semantics isn't compatible with compositional semantics and the denial of an analytic/synthetic distinction, as full conceptual roles aren't compositional, and there's no way to specify a relevant subset.

Harman, G. 1974. Meaning and semantics. In (M. Munitz & P. Unger, eds) *Semantics and Philosophy*. New York University Press.

Harman, G. 1975. Language, thought, and communication. In (K. Gunderson, ed) *Language, Mind, and Knowledge*. University of Minnesota Press.

Harman, G. 1982. Conceptual role semantics. *Notre Dame Journal of Formal Logic* 28:242-56.

Meaning and content is determined by the role of symbols in thought (e.g. inference and perception). With remarks on indeterminacy, context-dependence, the linguistic division of labor, qualia, speech acts, and more.

Horowitz, A. 1992. Functional role and intentionality. *Theoria* 58:197-218.

Loar, B. 1982. Conceptual role and truth conditions. *Notre Dame Journal of Formal Logic* 23:272-83.

On the relation between conceptual role and truth-conditions. Contra Harman, truth-conditions are to an extent independent of conceptual role.

Loewer B. 1982. The role of 'Conceptual role semantics'. *Notre Dame Journal of Formal Logic* 23:305-15.

Contra Harman 1982, truth-conditions are central to a semantic theory.

Perlman, M. 1997. The trouble with two-factor conceptual role theories. *Minds and Machines* 7:495-513.

Silverberg, A. 1992. Putnam on functionalism. *Philosophical Studies* 67:111-31.

Argues against Putnam 1987 that conceptual role plays an important role in determining meaning. Appeals to the induction theory of Holland et al.

Toribio, J. 1997. Twin pleas: Probing content and compositionality. *Philosophy and Phenomenological Research* 57:871-89.

Warfield, T. A. 1993. On a semantic argument against conceptual role semantics. *Analysis* 53:298-304.

Contra Fodor and Lepore, meanings can be compositional even if inferential roles are not, as long as meanings only supervene on inferential role.

2.3f

Theories of Content, Misc

Bestor, T. W. 1991. Naturalizing semantics: New insights or old folly? *Inquiry* 34:285-310.

Brook, A. & Stainton, R. 1997. Fodor's new theory of content and computation. *Mind and Language* 12:459-74.

Callaway, H. G. 1995. Intentionality naturalized: Continuity, reconstruction, and instrumentalism. *Dialectica* 49:147-68.

Churchland, P. M. & Churchland, P. S. 1983. Stalking the wild epistemic engine. *Nous* 17:5-18.

On "translational" (conceptual) and "calibrational" (referential) content. Relation of content issues to computational issues.

Cummins, R. 1989. *Meaning and Mental Representation*. MIT Press.

Critiques other views, offers interpretational semantics.

Cummins, R. 1996. *Representations, Targets, and Attitudes*. MIT Press.

Dennett, D. C. 1991. Ways of establishing harmony. In (B. McLaughlin, ed) *Dretske and his Critics*. Blackwell.

On the ways in which meanings can come to cohere with their causal roles: learning, natural selection, and design. Criticizes Dretske for undervaluing the latter two: all three are in the same boat.

Dretske, F. 1986. Aspects of cognitive representation. In (M. Brand & R. Harnish, eds) *The Representation of Knowledge and Belief*. University of Arizona Press.

On the reference and content of representations. Reference is determined by causation; content, i.e. representation "as", is determined by functional role, when functioning normally in natural habitat.

Shapiro, L. A. 1997. The nature of nature: Rethinking naturalistic theories of intentionality. *Philosophical Psychology* 10:309-322.

Silvers, S. 1991. On naturalizing the semantics of mental representation. *British Journal for the Philosophy of Science* 42:49-73.

Stalnaker, R. 1991. How to do semantics for the language of thought. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and His Critics*. Blackwell.

On some tensions in Fodor's view of content: e.g. narrow content must be dependent on functional role, which seems to lead to holism. The role of denotational semantics as a defense is unclear.

2.4

Representation (General) [see also 4.2]

Bickhard, M. 1993. Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence* 5:285-33.

Blachowicz, J. 1997. Analog representation beyond mental imagery. *Journal of Philosophy* 94:55-84.

Chomsky, N. 1980. Rules and representations. *Behavioral and Brain Sciences* 3:1-61.

Clark, A. 1995. Moving minds: Situating content in the service of real-time success. *Philosophical Perspectives* 9:89-104.

Dalenoort, G. J. 1990. Toward a general theory of representation. *Psychological Research* 52:229-237.

Fodor, J. A. 1986. Why paramecia don't have mental representations. *Midwest Studies in Philosophy* 10:3-23.

Because paramecia can't respond to non-nomic properties of the stimulus. Perceptual categories vs. sensory manifolds.

Freeman, W. & Skarda, C. A. 1990. Representations: who needs them? In (J. McGaugh, J. Weinberger, & G. Lynch) *Brain Organization and Memory*. Guilford Press.

Gillett, G. 1989. Representations and cognitive science. *Inquiry* 32:261-77.

Goldman, A. 1986. Constraints on representation. In (M. Brand & R. Harnish, eds) *The Representation of Knowledge and Belief*. University of Arizona Press.

Grush, R. 1997. The architecture of representation. *Philosophical Psychology* 10:5-23.

Hatfield, G. 1989. Computation, representation and content in noncognitive theories of perception. In (S. Silvers, ed) *ReRepresentation*. Kluwer.

Hogan, M. 1994. What is wrong with an atomistic account of mental representation. *Synthese* 100:307-27.

Jackendoff, R. 1991. The problem of reality. *Nous* 25:411-33.

On the philosophical (inward-out) vs. psychological (outward-in) approaches to the mind-world relation; the psychological approach is more useful in understanding representation. Internal reality is an imperfect construction.

Kukla, R. 1992. Cognitive models and representation. *British Journal for the Philosophy of Science* 43:219-32.

Lloyd, D. 1987. Mental representation from the bottom up. *Synthese* 70:23-78.

Lycan, W. G. 1989. Ideas of representation. In (Weissbord, ed) *Mind, Value and Culture: Essays in Honor of E. M. Adams*. Ridgeview.

Matthews, R. J. 1984. Troubles with representationalism. *Social Research* 51:1065-97.

Millikan, R. G. 1995. Pushmi-pullyu representations. *Philosophical Perspectives* 9:185-200.

Richardson, R. C. 1981. Internal representation: Prologue to a theory of intentionality. *Philosophical Topics* 12:171-212.

Shanon, B. 1991. Representations -- senses and reasons. *Philosophical Psychology* 4:355-74.

On different senses of "representation" -- external, experiential, mental locus, substrate of meaning, mediating functions, technicalpsychological.

Shanon, B. 1993. *The Representational and the Presentational: An Essay on Cognition and the Study of Mind*. Prentice-Hall.

Sober, E. 1976. Mental representations. *Synthese* 33:101-48.

Sterelny, K. 1995. Basic minds. *Philosophical Perspectives* 9:251-70.

van Gulick, R. 1982. Mental representation: A functionalist view. *Pacific Philosophical Quarterly* 63:3-20.

On the distinction between representation and representation-use.

Wallis, C. 1994. Representation and the imperfect ideal. *Philosophy of Science* 61:407-28.

2.5 The Explanatory Role of Content (Dretske, etc)

Adams, F. 1991. Causal contents. In (B. McLaughlin, ed) *Dretske and his Critics*. Blackwell.

On Dretske's account of the causal role of content. Addresses some objections: Dennett's worries about intrinsic intentionality, Fodor's about external causal powers, and some worries about syntax.

Baker, L. R. 1991. Dretske on the explanatory role of belief. *Philosophical Studies* 63:99-111.

Bogdan, R. J. 1989. Does semantics run the psyche? *Philosophy and Phenomenological Research* 49:687-700.

A critique of Fodor. Semantics per se doesn't cause. Also, Fodor's is an account of the what, not the how, of semantics. Somewhat bizarre.

Cummins, R. 1991. Mental meaning in psychological explanation. In (B. McLaughlin, ed) *Dretske and his Critics*. Blackwell.

Criticizes Dretske's account of the role of content, especially because of its dependence on an organism's history; also, it may not cohere with work in cognitive science. Argues for an interpretational, not a causal account.

Devitt, M. 1991. Why Fodor can't have it both ways. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Dretske, F. 1987. The explanatory role of content. In (R. Grimm & D. Merrill, eds) *Contents of Thought*. University of Arizona Press.

Content must explain why (not how) an internal state caused a certain output. The explanation is given in terms of what a state has historically indicated. With thermostats and sea-snails as examples. Comments by Cummins, and reply.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Dretske, F. 1990. Does meaning matter? In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Dretske, F. 1994. Reply to Slater and Garcia-Carpintero. *Mind and Language* 9:203-8.

Dretske, F. 1995. Reply: Causal relevance and explanatory exclusion. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Dretske, F. 1996. The explanatory role of content: Reply to Melnyk and Noordhof. *Mind and Language* 11:223-29.

Elder, C. L. 1996. Content and the subtle extensionality of "... explains ...". *Philosophical Quarterly* 46:320-32.

Fodor, J. A. 1986. Banish DisContent. In (J. Butterfield, ed) *Language, Mind, and Logic*. Cambridge University Press. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Garcia-Carpintero, M. 1994. Dretske on the causal efficacy of meaning. *Mind and Language* 9:181-202.

Hassrick, B. 1995. Fred Dretske on the explanatory role of semantic content. *Conference* 6:59-66.

Horgan, T. 1991. Actions, reasons, and the explanatory role of content. In (B. McLaughlin, ed) *Dretske and his Critics*. Blackwell.

Distinguishes three problems of mental causation (extrinsic factors, exclusion of the nonphysical, anomalism). Criticizes Dretske's theory (can't handle unlearned or here-and-now reasons), offers a counterfactual account.

Melnyk, A. 1996. The prospects for Dretske's account of the explanatory role of belief. *Mind and Language* 11:203-15.

Noordhof, P. 1996. Accidental associations, local potency, and a dilemma for Dretske. *Mind and Language* 11:216-22.

Perry, J. & Israel, D. 1991. Fodor and psychological explanation. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Pylyshyn, Z. W. 1987. What's in a mind? *Synthese* 70:97-122.

We must individuate mental states by semantics, not just by function, as we need representation to capture generalizations about behavior; particularly due to the information-sensitivity and stimulus-independence of behavior.

Slater, C. 1994. Discrimination without indication: Why Dretske can't lean on learning. *Mind and Language* 9:163-80.

Wallis, C. 1994. Using representation to explain. In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

2.6 Concepts

Barber, A. 1998. The pleonasticity of talk about concepts. *Philosophical Studies* 89:53-86.

Brown, H. 1986. Sellars, concepts, and conceptual change. *Synthese* 68:275-307.

Burge, T. 1993. Concepts, definitions, and meaning. *Metaphilosophy* 24:309-25.

Churchland, P. M. 1998. Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy* 95:5-32.

Cussins, A. 1990. The connectionist construction of concepts. In (M. Boden, ed) *The Philosophy of AI*. Oxford University Press.

Connectionism builds up concepts from the nonconceptual level. From nonconceptual content (e.g. perceptual experiences) to the emergence of objectivity.

- Fodor, J. & Lepore, E. 1996. The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition* 58:253-70.
- Fodor, J. 1995. Concepts: A potboiler. *Cognition* 50:133-51. Also in (E. Villanueva, ed) *Content*. Ridgeview.
- Fodor, J. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.
- Franks, B. 1992. Realism and folk psychology in the ascription of concepts. *Philosophical Psychology* 5:369-90.
- Gardenfors, P. 1997. Meanings as conceptual structures. In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.
- Gauker, C. 1993. An extraterrestrial perspective on conceptual development. *Mind and Language* 8:105-30.
- Grandy, R. E. 1989. Concepts, prototypes, and information. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.
- Jackendoff, R. 1989. What is a concept, that a person may grasp it? *Mind and Language* 4:68-102.
- Khalidi, M. A. 1995. Two concepts of concept. *Mind and Language* 10:402-22.
- Levine, A. & Bickhard, MH. 1999. Concepts: Where Fodor went wrong. *Philosophical Psychology* 12:5-23.
- Livingston, K. R. 1989. Concepts, categories, and epistemology. *Philosophia* 19:265-300.
- Neisser, U. (ed) 1981. *Concepts and Conceptual Development*. Cambridge University Press.
- Osherson, D. N. & Smith, E. E. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition* 9:35-58.
- Margolis, E. 1995. The significance of the theory analogy in the psychological study of concepts. *Mind and Language* 10:45-71.
- Margolis, E. 1998. How to acquire a concept. *Mind and Language* 13:347-369.
- Margolis, E. 1999. What is conceptual glue? *Minds and Machines* 9:241-255.

- Margolis, E. & Laurence, S. 1999. *Concepts: Core Readings*. MIT Press.
- Millikan, R. G. 1994. On unclear and indistinct ideas. *Philosophical Perspectives* 8:75-100.
- Millikan, R. G. 1997. A common structure for concepts of individuals, stuffs, and kinds: More mama, more milk, and more mouse. *Behavioral and Brain Sciences*.
- Peacocke, C. 1989. What are concepts? *Midwest Studies of Philosophy* 14.
- Peacocke, C. 1989. Possession conditions: A focal point for theories of concepts. *Mind and Language* 4:51-56.
- Peacocke, C. 1991. The metaphysics of concepts. *Mind* 100:525-46.
- Peacocke, C. 1992. *A Study of Concepts*. MIT Press.
- Peacocke, C. 1996. *Precis of A Study of Concepts*. *Philosophy and Phenomenological Research* 56:407-52.
- A symposium on the book, with comments by Heal, Rey, Papineau.
- Peacocke, C. 1996. Can a theory of concepts explain the a priori: A reply to Skorupski. *International Journal of Philosophical Studies* 4:154-60.
- Peacocke, C. 1996. The relation between philosophical and psychological theories of concepts. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.
- Peacocke, C. 2000. Theories of concepts: A wider task. *European Journal of Philosophy* 8:298-321.
- Perlman, M. 2000. *Conceptual Flux: Mental Representation, Misrepresentation, and Concept Change*. Kluwer.
- Pitt, D. 1999. In defense of definitions. *Philosophical Psychology* 12:139-156.
- Ramsey, W. 1992. Prototypes and conceptual analysis. *Topoi* 11:59-70.
- On the significance of psychological work on concepts for philosophical conceptual analysis -- simple, precise analyses do not exist in general.
- Rey, G. 1983. Concepts and stereotypes. *Cognition* 15:237-62.
- Rips, L. J. 1995. The current status of research on concept combination. *Mind and Language* 10:72-104.

Sellars, W. 1948. Concepts as involving laws and inconceivable without them. *Philosophy of Science* 15:287-313.

Sellars, W. 1974. Conceptual change. In *Essays in Philosophy and its History*. Reidel.

Smith, E. E. & Medin, D. L. 1981. *Categories and Concepts*. Harvard University Press.

Sosa, E. 1993. Abilities, concepts, and externalism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

On concepts as abilities, and on construals of abilities that lead to internalism and externalism. Maybe the relevant abilities are characterized externally but determined internally. Remarks on Putnam, Davidson, Burge.

Thagard, P. 1990. Concepts and conceptual change. *Synthese* 82:255-74.

van Brakel, J. 1991. Meaning, prototypes, and the future of cognitive science. *Minds and Machines* 1:233-57.

Vision, G. 2001. Flash! Fodor splits the atom. *Analysis* 61:5-10.

Weitz, M. 1988. *Theories of Concepts: A History of the Major Philosophical Traditions*. Routledge.

Woodfield, A. 1991. Conceptions. *Mind* 100:547-72.

2.7

Meaning Holism

Abbott, B. 2000. Fodor and Lepore on meaning similarity and compositionality. *Journal of Philosophy* 97:454-6.

Becker, K. 1998. On the perfectly general nature of instability in meaning holism. *Journal of Philosophy* 95:635-640.

Bilgrami, A. 1998. Why holism is harmless and necessary. *Philosophical Perspectives* 12:105-126.

Block, N. 1995. An argument for holism. *Proceedings of the Aristotelian Society* 95:151-70.

Uses Putnam's "Ruritania" example to argue that narrow content, if it exists, is holistic. Twins in different

communities start with same narrow content, diverge by acquiring new beliefs; so belief change affects narrow content.

Callaway, H. G. 1992. Meaning holism and semantic realism. *Dialectica* 46:41-59.

Churchland, P. M. 1998. Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal Of Philosophy* 95:5-32.

Devitt, M. 1994. A critique of the case for semantic holism. *Philosophical Perspectives* 8:281-306.

Devitt, M. 1994. Semantic localism: Who needs a principled basis? In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*. Holder-Pichler-Tempsky.

Esfeld, M. 1998. Holism and analytic philosophy. *Mind* 107:365-80.

Fodor, J. A. & LePore, E. 1992. *Holism: A Shopper's Guide*. Blackwell.

Rebutting arguments for meaning holism: those based on confirmation holism (Quine), normativity of interpretation (Davidson, Dennett, Lewis), and functional-role semantics (Block, Field, Churchland).

Fodor, J. A. & LePore, E. 1993. *Precis of Holism: A Shopper's Guide*. *Philosophy and Phenomenological Research* 53:637-682.

A discussion of *Holism* with comments by Devitt, Rey, McLaughlin, Brandom, and Churchland, and a reply by Fodor and Lepore.

Gauker, C. 1993. Holism without meaning: A critical review of Fodor and Lepore's *Holism: A Shopper's Guide*. *Philosophical Psychology* 6:441-49.

Harrell, M. 1996. Confirmation holism and semantic holism. *Synthese* 109:63-101.

Heal, J. 1994. Semantic holism: Still a good buy. *Proceedings of the Aristotelian Society* 68:325-39.

A critique of Fodor and Lepore. Distinguishes versions of holism, and argues for a weak version. Real thinkers are subjects, which imposes constraints on the interrelations of thoughts. Science fiction is irrelevant here.

Kukla, A. & Kukla, R. 1989. Meaning holism and intentional psychology. *Analysis* 173-53.

Contra Fodor, meaning holism is compatible with intentional psychology. Most psychological generalizations quantify over contents, rather than appealing to specific contents.

Lormand, E. 1996. How to be a meaning holist. *Journal of Philosophy* 93:51-73.

Margolis, E. & Laurence, S. 1998. Multiple meanings and stability of content. *Journal of Philosophy* 5:255-63.

McClamrock, R. 1989. Holism without tears: Local and global effects in cognitive processing. *Philosophy of Science* 56:258-74.

Miller, R. B. 1997. One bad and one not very good argument against holism. *Australasian Journal of Philosophy* 75:234-40.

A nice criticism of Fodor and Lepore's arguments that holism implies (a) the nonexistence of intentional laws and (b) the nonlearnability of language.

Pagin, P. 1997. Is compositionality compatible with holism? *Mind and Language* 12:11-33.

Perry, J. 1994. Fodor and Lepore on holism. *Philosophical Studies* 73:123-58.

The argument from atomism and the failure of the analytic/synthetic distinction to holism fails. On the many different interpretations of holism and atomism: there is a reasonable molecularist position.

Senor, T. D. 1992. Two-factor theories, meaning holism, and intentionalistic psychology: A reply to Fodor. *Philosophical Psychology* 5:133-51.

Silverberg, A. 1994. Meaning holism and intentional content. *Pacific Philosophical Quarterly* 75:29-53.

Stich, S. P. 1983. Some evidence against narrow causal theories of belief. In *From Folk Psychology to Cognitive Science*. MIT Press.

Talmage, C. J. L. & Mercer, M. 1991. Meaning holism and interpretability. *Philosophical Quarterly* 41:301-15.

Talmage, C. J. L. 1998. Semantic localism and the locality of content. *Erkenntnis* 48:101.

2.8

Mental Content, Misc

Allen, C. 1992. Mental content. *British Journal for the Philosophy of Science* 43:537-53.

Beckermann, A. 1996. Is there a problem about intentionality? *Erkenntnis* 45:1-24.

- Bilgrami, A. 1992. *Belief and Meaning: The Unity and Locality of Mental Content*. Blackwell.
- Bogdan, R. J. 1986. The manufacture of belief. In (R. Bogdan, ed) *Belief: Form, Content, and Function*. Oxford University Press.
- Butler, K. 1995. Content, context, and compositionality. *Mind and Language* 10:3-24.
- Cummins, R. 1991. Form, interpretation, and the uniqueness of content: A response to Morris. *Minds and Machines* 1:31-42.
- Morris 1991 is wrong: formal individuation is easy, and objectively determinate content isn't needed. External grounding is also irrelevant.
- Devitt, M. 1994. The methodology of naturalistic semantics. *Journal of Philosophy* 91:519-44.
- Engel, P. 2000. Wherein lies the normative dimension in meaning and mental content? *Philosophical Studies* 100:305-321.
- Fodor, J. A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.
- Fodor, J. A. 1990. *A Theory of Content and Other Essays*. MIT Press.
- Fodor, J. A. 1994. *The Elm and the Expert*. MIT Press.
- Garfield, J. L. 2000. The meanings of "meaning" and "meaning": Dimensions of the sciences of mind. *Philosophical Psychology* 31:421-440.
- Gillett, G. 1992. *Representation, Meaning, and Thought*. Oxford University Press.
- Haldane, J. J. 1989. Naturalism and the problem of intentionality. *Inquiry* 32:305-22.
- Haugeland, J. 1990. The Intentionality All-Stars. *Philosophical Perspectives* 4:383-427.
- Intentionality around the diamond: neoCartesianism, neobehaviorism, neopragmatism.
1B=Fodor/Pylyshyn, 2B=Dennett/Quine, 3B=Heidegger/Sellars. SS=Wittgenstein. RF=Searle,
CF=Skinner, LF=Rorty/Derrida. Lots of fun.
- Horgan, T. 1994. Naturalism and intentionality. *Philosophical Studies* 76:301-26.
- Jacob, P. 1997. *What Minds Can Do: Intentionality in a Non-intentional World*. Cambridge University Press.

Kaye, L. J. 1995. A scientific psychologicistic foundation for theories of meaning. *Minds and Machines* 5:187-206.

Lehrer, K. 1986. Metamind: Belief, consciousness and intentionality. In (R. Bogdan, ed) *Belief: Form, Content, and Function*. Oxford University Press.

Lycan, W. G. 1986. Thoughts about things. In (M. Brand & R. Harnish, ed) *The Representation of Knowledge and Belief*. University of Arizona Press.

Madell, G. 1989. Physicalism and the content of thought. *Inquiry* 32:107-21.

Maloney, J. C. 1989. *The Mundane Matter of the Mental Language*. Cambridge University Press.

Martin, C. B. & Pfeifer, K. 1986. Intentionality and the non-psychological. *Philosophy and Phenomenological Research* 46:531-54.

McDowell, J. 1998. Lecture III: Intentionality as a relation. *Journal Of Philosophy* 95:471-491.

McGinn, C. 1989. *Mental Content*. Blackwell.

McManus, D. 2000. Boghossian, Miller and Lewis on dispositional theories of meaning. *Mind and Language* 15:393-399. Miller, A. 1997. Boghossian on reductive dispositionalism about conten: The case strengthened. *Mind and Language* 12:1-10.

Morris, M. 1991. Why there are no mental representations. *Minds and Machines* 1:1-30.

There can be no non-stipulative content to non-semantically individuated tokens. Mostly a critique of Cummins; also Fodor and Dennett.

Newton, N. 1996. *Foundations of Understanding*. John Benjamins.

Pacherie, E. 2000. The content of intentions. *Mind and Language* 15:400-432.

Peacocke, C. 1986. *Thoughts: An Essay on Content*. Blackwell.

Peacocke, C. 1991. Content and norms in a natural world. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Pollock, J. 1990. Understanding the language of thought. *Philosophical Studies* 58:95-120.

Remarks on a number of aspects of mental content -- narrow, propositional, qualitative -- with respect to functionalism and the language of thought. With comments by Baker.

- Priest, G. 2000. Objects of thought. *Australasian Journal of Philosophy* 78:494-502.
- Prinz, J. 2000. The duality of content. *Philosophical Studies* 100:1-34.
- Schiffer, S. 1981. Truth and the theory of content. In (H. Parret, ed) *Meaning and Understanding*. Berlin.
- Schiffer, S. 1987. *Remnants of Meaning*. MIT Press.
- Sellars, W. & Chisholm, R. 1957. Intentionality and the mental: A correspondence. *Minnesota Studies in the Philosophy of Science* 2:507-39.
- Stalnaker, R. 1999. *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press.
- Sterelny, K. 1990. *The Representational Theory of Mind*. Blackwell.
- Stich, S. P. 1982. On the ascription of content. In (A. Woodfield, ed) *Thought and Object*. Oxford University Press.
- On the tacit theories underlying the folk psychology of belief: beliefs are states associated with typical causal patterns. With remarks on ambiguity, content identity and similarity, and environmental dependence.
- Stich, S. P. , and Laurence, S. 1994. Intentionality and naturalism. *Midwest Studies in Philosophy* 19:159-82. Reprinted in (Stich) *Deconstructing the Mind*. Oxford University Press, 1996.
- Argues that a failure to "naturalize" intentionality won't lead to disasters such as irrealism, irrelevance, or non-science, whether naturalization is understood as analysis, property identity, supervenience, or whatever.
- Tye, M. 1994. Naturalism and the problem of intentionality. *Midwest Studies in Philosophy* 19:122-42.
- There's no deep problem of naturalism about intentionality, as we know it's true already. The real puzzle is that of finding a mechanism to close the gap, e.g. via analysis or essentialism. But naturalism doesn't require that.

Part 3: Metaphysics of Mind [1026]

Part of [Contemporary Philosophy of Mind: An Annotated Bibliography](#).

Compiled by [David J. Chalmers](#), Department of Philosophy, University of Arizona, Tucson AZ 85721. E-mail: chalmers@arizona.edu.

- 3.1 [Supervenience](#) [94]
 - 3.1a [Psychophysical Supervenience](#) (Kim, etc) [14]
 - 3.1b [Supervenience and Physicalism](#) [25]
 - 3.1c [Technical Issues in Supervenience](#) [27]
 - 3.1d [Supervenience, General](#) [28]
- 3.2 [Reduction](#) [82]
 - 3.2a [Multiple Realizability and Reduction](#) [25]
 - 3.2b [Nonreductive Materialism](#) [28]
 - 3.2c [Reduction in Psychology](#) [14]
 - 3.2d [Miscellaneous](#) [15]
- 3.3 [Other Psychophysical Relations](#) [127]
 - 3.3a [Physicalism](#) [29]
 - 3.3b [Token Identity](#) [7]
 - 3.3c [Emergence](#) [58]
 - 3.3d [Dualism](#) [21]
 - 3.3e [Miscellaneous](#) [12]
- 3.4 [Functionalism](#) [63]
 - 3.4a [Causal Role Functionalism](#) (Armstrong/Lewis) [17]
 - 3.4b [Machine Functionalism](#) (Putnam) [15]
 - 3.4c [Miscellaneous](#) [31]
- 3.5 [Other Psychophysical Theories](#) [238]
 - 3.5a [Logical Behaviorism](#) (Ryle, etc) [53]
 - 3.5b [Identity Theory](#) (Smart, Place, etc) [98]
 - 3.5c [Eliminative Materialism](#) (Rorty, Feyerabend) [30]
 - 3.5d [Anomalous Monism](#) (Davidson) [57]
- 3.6 [Mental Causation](#) [58]
- 3.7 [Personal Identity](#) [189]
 - 3.7a [Personal Identity, General](#) [76]
 - 3.7b [Parfit on Personal Identity](#) [49]
 - 3.7c [Persons](#) [23]
 - 3.7d [Split Brains](#) [22]

- 3.7e [Multiple Personality Disorder](#) [19]
- 3.8 [Free Will](#) [129]
- 3.9 [The Problem of Other Minds](#) [46]

3.1 Supervenience

3.1a

Psychophysical Supervenience (Kim, etc)

Crane, T. 1991. Why indeed? Papineau on supervenience. *Analysis* 51:32-7.

Contra Papineau 1989: the assumption of completeness is false or trivial. Maybe the mental is part of a complete physics. With response by Papineau.

Elugardo, R. 1988. Against weak psychophysical supervenience. *Dialectica* 42:129-43.

Various objections to Kim's arguments for supervenience. Not all internal states relevant to I/O relations are psychological states. Strange.

Kim, J. 1979. Causality, identity and supervenience in the mind-body problem. *Midwest Studies in Philosophy* 4:31-49.

Supervenience of the mental on the physical is what is required to make mental causation possible. Very nice.

Kim, J. 1982. Psychophysical supervenience. *Philosophical Studies* 41:51-70. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Internal mental states (i.e. ones that are not rooted outside) supervene on synchronous internal physical states, and internal states are all that is relevant in the explanation of behavior.

Kim, J. 1982. Psychophysical supervenience as a mind-body theory. *Cognition and Brain Theory* 5:129-47.

Distinguishes weak (within-world) vs strong (across-worlds) supervenience. Relates to reduction, internal/external mental states, and various theories.

Kim, J. 1997. Supervenience, emergence, and realization in the philosophy of mind. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Lewis, H. 1985. Is the mental supervenient on the physical? In (B. Vermazen & M. Hintikka, eds) *Essays*

on Davidson. Oxford University Press.

On some problems with supervenience, the relation between supervenience and reduction, and on reasons for accepting psychophysical supervenience. Loose.

Loar, B. 1993. Can we confirm supervenient properties? In (E. Villanueva, ed) *Naturalism and Normativity*. Ridgeview.

If mental properties are supervenient but irreducible to physical/functional properties, we can't confirm them. Confirmation requires an indispensable explanatory role, which irreducibility precludes. With comments by Schiffer.

Macdonald, C. 1995. Psychophysical supervenience, dependency, and reduction. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Merricks, T. 1998. Against the doctrine of microphysical supervenience. *Mind* 107: 59-72.

Noonan, H. 1999. Microphysical supervenience and consciousness. *Mind* 108:755-9.

Papineau, D. 1989. Why supervenience? *Analysis* 50:66-71.

Psychophysical supervenience follows from completeness of physical laws.

Papineau, D. 1995. Arguments for supervenience and physical realization. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Witmer, D. G. 1998. What is wrong with the manifestability argument for supervenience? *Australasian Journal of Philosophy* 76:84-89.

3.1b Supervenience and Physicalism [see also 3.3a]

Armstrong, D. M. 1982. Metaphysics and supervenience. *Critica* 42:3-17.

Argues that everything is logically supervenient on the physical. Considers classes, possibilities, numbers, universals, and objects of thought.

Bailey, A. 1998. Supervenience and physicalism. *Synthese* 117:53-73.

Chalmers, D. J. 1996. Supervenience and materialism. In *The Conscious Mind* (pp. 41-42). Oxford University Press, 1996.

Charles, D. 1992. Supervenience, composition, and physicalism. In (D. Charles & K. Lennon, eds)

Reduction, Explanation and Realism. Oxford University Press.

Francescotti, R. M. 1998. Defining "physicalism". *Journal of Mind and Behavior* 19:51-64.

Haugeland, J. 1984. Ontological supervenience. *Southern Journal of Philosophy Supplement* 22:1-12.

Supervenience is all we need for materialism. Various materialist arguments (unity, "nothing but", history, fear of darkness, simplicity, law) don't support physical exhaustion & token identity, over and above supervenience.

Hellman, G. & Thomson, F. 1975. Physicalism: ontology, determination and reduction. *Journal of Philosophy* 72:551-64.

Hellman, G. & Thomson, F. 1977. Physicalist materialism. *Nous* 11:309-45.

Some applications of the earlier treatment: examples of determination without reduction; the status of properties and universals; the mental; the life sciences; modalities and essentialism; theoretical equivalence.

Hellman, G. 1985. Determination and logical truth. *Journal of Philosophy* 82:607-16.

Some remarks on determination, physicalism, model theory, and logical truth.

Horgan, T. 1981. Token physicalism, supervenience, and the generality of physics. *Synthese* 49:395-413.

Argues that the generality of physics should be a supervenience thesis, not token physicalism. Fodor's token physicalism is untenable but might be saved with an appropriate view of events.

Horgan, T. 1982. Supervenience and microphysics. *Pacific Philosophical Quarterly* 63:29-43.

An account of how all facts supervene on microphysical facts, and how all intrinsic facts supervene on intrinsic microphysical facts.

Horgan, T. 1984. Supervenience and cosmic hermeneutics. *Southern Journal of Philosophy Supplement* 22:19-38.

Laplacean demon's job: number crunching, plus cosmic hermeneutics to explain high-level truths. All high-level truths follow from low-level by meaning constraints. Application to theoretical/mentalist/everyday terms. Nice.

Jack, A. 1994. Materialism and supervenience. *Australasian Journal of Philosophy* 72:426-43.

Supervenience is neither necessary nor sufficient for materialism. With various (contentious) counterexamples. So we need a different formulation.

Kirk, R. 1996. Strict implication, supervenience, and physicalism. *Australasian Journal of Philosophy* 74:244-57.

Argues for strict implication rather than supervenience as a formulation of "minimal physicalism" (unless supervenience is formulated just right).

Lewis, D. 1983. New work for a theory of universals. *Australasian Journal of Philosophy*.

Formulates a definition of materialism: among worlds where no natural properties alien to our worlds are instantiated, no two differ without differing physically. With a lot of other material on universals.

Melnyk, A. 1991. Physicalism: From supervenience to elimination. *Philosophy and Phenomenological Research* 51:573-87.

How can supervenience, as a relationship between ontologically distinct properties, be explained? Modal realism and grand-properties don't work. Eliminativism about supervenient properties is the only possibility.

Moreland, J. P. 1999. Should a naturalist be a supervenient physicalist? *Metaphilosophy* 29:35-57.

Moser, P, & Trout, J. D. 1996. Physicalism, supervenience, and dependence. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Pettit, P. 1993. A definition of physicalism. *Analysis* 53:213-23.

Defines physicalism in terms of claims that microphysical entities constitute everything and that microphysical laws govern everything. With a reply by Crane.

Rowlands, M. 1995. *Supervenience and Materialism*. Avebury.

Seager, W. E. 1988. Weak supervenience and materialism. *Philosophy and Phenomenological Research* 48:697-709.

Weak supervenience provides a more tenable form of materialism than strong supervenience, because of inverted spectrum possibilities, etc.

Snowdon, P. F. 1989. On formulating materialism and dualism. In (J. Heil, ed) *Cause, Mind, and Reality: Essays Honoring C. B. Martin*. Kluwer.

A construal of materialism in terms of constitution, not identity. Discusses the entailment between physical properties and mental properties; considers a nonreductive physicalism and a primitive dualism.

Sturgeon, S. 1998. Physicalism and overdetermination. *Mind* 107:411-32.

Wilson, J. 1999. How superduper does a physicalist supervenience need to be? *Philosophical Quarterly* 49:33-52.

Witmer, D. G. 1999. Supervenience physicalism and the problem of extras. *Southern Journal of Philosophy* 37:315-31.

3.1c Technical Issues in Supervenience

Bacon, J. 1986. Supervenience, necessary coextensions, and reducibility. *Philosophical Studies* 49:163-76.

A modal-logic analysis of the relations between various notions of supervenience. Most concepts of supervenience entail necessary co-extension, under certain closure assumptions for properties.

Bacon, J. 1995. Weak supervenience supervenes. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Bonevac, D. 1988. Supervenience and ontology. *American Philosophical Quarterly* 25:37-47.

A model-theoretic treatment of supervenience, in terms of relations between theories. Supervenience turns out to be equivalent to reduction.

Bovens, L. 1994. Principles of supervenience. *Australasian Journal of Philosophy* 72:294-301.

Divers, J. 1996. Supervenience for operators. *Synthese* 106:103-12.

Forrest, P. 1988. Supervenience: The grand-property hypothesis. *Australasian Journal of Philosophy* 66:1-12.

A nonreductive supervenience hypothesis: supervenient properties are properties of properties, e.g intrinsic goodness is a property of an object's nature.

Forrest, P. 1992. Universals and universalisability: An interpretation of Oddie's discussion of supervenience. *Australasian Journal of Philosophy* 70:93-98.

Grim, P. 1997. Worlds by supervenience: Some further problems. *Analysis* 2:146-51.

Grimes, T. 1991. Supervenience, determination, and dependency. *Philosophical Studies* 62:81-92.

On dependency supervenience (B properties determine A properties) versus determination supervenience (A properties need B properties).

Grimes, T. 1995. The Tweedledum and Tweedledee of supervenience. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Humberstone, I. L. 1992. Some structural and logical aspects of the notion of supervenience. *Logical Analysis* 35:101-37.

Kim, J. 1984. Concepts of supervenience. *Philosophy and Phenomenological Research* 45:153-76. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Distinguishes weak and strong supervenience. A mistaken proof that strong and global supervenience are equivalent. Strong supervenience implies a kind of reduction, but not an explanatorily useful reduction.

Kim, J. 1987. 'Strong' and 'global' supervenience revisited. *Philosophy and Phenomenological Research* 48:315-26. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Reasons why global supervenience doesn't entail strong supervenience, and trying to rescue global supervenience as a useful notion. Suggests a similarity-based notion of global supervenience.

Kim, J. 1988. Supervenience for multiple domains. *Philosophical Topics* 16:129-50. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

How properties in one domain can supervene on properties in another, with or without co-ordination between domains. Relation to global supervenience.

Klagge, J. C. 1995. Supervenience: Model theory or metaphysics? In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Marras, A. 1993. Supervenience and reducibility: An odd couple. *Philosophical Quarterly* 43:215-222.

Supervenience doesn't entail reducibility, as necessary coextension doesn't suffice, and is incompatible with reducibility, due to ontological asymmetry.

McLaughlin, B. P. 1995. Varieties of supervenience. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Distinguishes modal-operator and possible-worlds versions of supervenience, and explicates global supervenience and its relation to weak and strong. With remarks on multiple-domain supervenience and the relation to reduction.

McLaughlin, B. P. 1997. Supervenience, vagueness, and determination. *Philosophical Perspectives* 11:209-30.

Melnyk, A. 1997. On the metaphysical utility of claims of global supervenience. *Philosophical Studies*

87:277-308.

Moser, P. K. 1992. Physicalism and global supervenience. *Southern Journal of Philosophy* 30:71-82.

Argues that global supervenience has epistemological problems -- how could we ever know that it holds, and that certain worlds are impossible?

Oddie, G. & Tichy, P. 1990. Resplicing properties in the supervenience base. *Philosophical Studies* 58:259-69.

Closure under resplicing makes supervenience both too narrow and too wide. Weak supervenience is generally too weak to capture the dependence relation.

Oddie, G. 1991. Supervenience and higher-order universals. *Australasian Journal of Philosophy* 69:20-47.

Paull, R. C. & Sider, T. R. 1992. In defense of global supervenience. *Philosophy and Phenomenological Research* 52:833-53.

Gives a proof of the distinction between strong and global supervenience that improves on Petrie's, and argues contra Kim that global supervenience is a perfectly reasonable dependence relation for physicalism.

Petrie, B. 1987. Global supervenience and reduction. *Philosophy and Phenomenological Research* 48:119-30.

Defending global supervenience: it's weaker than strong supervenience, as base properties of other individuals are relevant. It doesn't entail type or token reducibility. On the relation to implicit definability and reduction.

Post, J. F. 1995. "Global" supervenient determination: Too permissive? In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Shagrir, O. 1999. More on global supervenience. *Philosophy and Phenomenological Research* 59:691-701.

van Cleve, J. 1990. Supervenience and closure. *Philosophical Studies* 58:225-38.

Properties in supervenience relations shouldn't be closed under negation or resplicing, due to bad consequences. With reply by Bacon on resplicing.

3.1d

Supervenience, General

Blackburn, S. 1984. Supervenience revisited. In (I. Hacking, ed) *Exercises in Analysis: Essays by Students of Casimir Lewy*. Cambridge University Press.

On the incompatibility of weak supervenience without strong supervenience and realism. With discussion of various strengths of necessity involved in supervenience claims, and application to moral realism and anomalous monism.

Currie, G. 1984. Individualism and global supervenience. *British Journal for the Philosophy of Science* 35:345-58.

How social facts supervene on the totality of individual facts. Application to belief, etc.

Enc, B. 1996. Nonreducible supervenient causation. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Grimes, T. 1988. The myth of supervenience. *Pacific Philosophical Quarterly* 69:152-60.

Supervenience is too weak to function as a dependency relation, as e.g. it can hold in two directions at once.

Hare, R. M. 1984. Supervenience. *Proceedings of the Aristotelian Society* 58:1-16.

On the universal conditionals that underlie supervenience, and the necessity thereof. A discussion of the necessity of moral, natural kind, and other sorts of supervenience. Contra Davidson, anomalous supervenience is silly.

Heil, J. 1995. Supervenience redux. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Hellman, G. 1992. Supervenience/determination a two-way street? Yes, but one of the ways is the wrong way! *Journal of Philosophy* 89:42-47.

Reply to Miller 1990. Miller underestimates the modal force of supervenience and invokes irrelevant dispositional properties.

Horgan, T. 1993. From supervenience to superdupervenience: Meeting the demands of a material world. *Mind* 102:555-86.

An overview of supervenience, with focus on the problem of explaining supervenience relations. With remarks on mental causation, emergence, physicalism, and reduction.

Kim, J. 1978. Supervenience and nomological incommensurables. *American Philosophical Quarterly* 15:149-56.

Developing and motivating the notion of supervenience. Investigating the relationship to reducibility and definability (equivalence, under certain conditions), and to microphysical determination.

Kim, J. 1984. Supervenience and supervenient causation. *Southern Journal of Philosophy Supplement* 22:45-56.

On weak/strong supervenience, and high-level causation via supervenience.

Kim, J. 1991. Supervenience as a philosophical concept. *Metaphilosophy* 21:1-27. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

A nice overview of supervenience and covariance.

Kim, J. 1993. *Supervenience and Mind*. Cambridge University Press.

A collection of articles on supervenience and causation in metaphysics and the philosophy of mind, with some added postscripts.

Kincaid, H. 1987. Supervenience doesn't entail reducibility. *Southern Journal of Philosophy* 25:343-56.

Supervenience doesn't entail reducibility, which is epistemological. The problem's not just huge disjuncts, but also the sharing of bases, no local correlations, and base-properties presupposing supervenient properties.

Kincaid, H. 1988. Supervenience and explanation. *Synthese* 77:251-81.

Argues that lower-level theories can explain supervenient but irreducible higher-level theories, but only under certain conditions, as low-level accounts don't have the relevant kind terms.

Klagge, J. C. 1988. Supervenience: Ontological and ascriptive. *Australasian Journal of Philosophy* 66:461-70.

On supervenience as an ontological relation (via metaphysical necessity) or as an ascriptive relation (via conceptual necessity). The first doesn't preclude the second. Moral realism and mental realism are in the same boat.

Loewer, B. 1995. An argument for strong supervenience. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

McLaughlin, B. P. 1983. Event supervenience and supervenient causation. *Southern Journal of Philosophy Supplement* 22:71-91.

McLaughlin, B. P. 1994. Varieties of supervenience. In (E. Savellos & O. Yalchin, eds) *Supervenience*:

New Essays.

On a number of issues: possible worlds vs modal notions, explicating global supervenience, the relation between weak/strong/global supervenience, multiple-domain supervenience, and implications for reduction.

Melnyk, A. 1997. On the metaphysical utility of claims of global supervenience. *Philosophical Studies* 87:277-308.

Merricks, T. 1998. Against the doctrine of microphysical supervenience. *Mind* 107:59-71.

Miller, R. B. 1990. Supervenience is a two-way street. *Journal of Philosophy* 87:695-701.

If supervening properties can make arbitrarily fine distinctions, then physical properties supervene on moral/aesthetic/mental properties.

Noonan, H. 1987. Supervenience. *Philosophical Quarterly* 37:78-85.

Contra Blackburn 1984 on the possibility of weak supervenience without strong supervenience, even with metaphysical necessity; using Nozick's concept structures, or indexical definitions. With application to moral realism.

Post, J. F. 1984. On the determinacy of valuation. *Philosophical Studies* 45:315-33.

Stalnaker, R. 1996. Varieties of supervenience. *Philosophical Perspectives* 10:221-42.

Distinguishes "reductionist" and "metaphysical" conceptions of supervenience. Also discusses the relation between strong and global supervenience, degrees of necessity, and the explanatory role of supervenience.

Teller, P. 1984. The poor man's guide to supervenience and determination. *Southern Journal of Philosophy Supplement* 22:137-62.

Compares the Hellman/Thompson notion of determination with Kim's development of supervenience. Uses these to investigate the concept of materialism, and argues that materialism isn't contingent.

Teller, P. 1985. Is supervenience just disguised reduction? *Southern Journal of Philosophy* 23:93-100.

van Brakel, J. 1996. Interdiscourse or supervenience relations: The primacy of the manifest image. *Synthese* 106:253-97.

Zangwill, N. 1997. Explaining supervenience: Moral and mental. *Journal of Philosophical Research* 22:509-18.

3.2

Reduction

3.2a

Reduction and Multiple Realizability

Antony, L. M. & Levine, J. 1997. Reduction with autonomy. *Philosophical Perspectives* 11:83-105. <
Bechtel, W. , & Mundale, J. 1999. Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science* 66:175-207.

Bickle, J. 1992. Multiple realizability and psychophysical reduction. *Behavior and Philosophy* 20:47-58.

Bickle, J. 1995. Connectionism, reduction, and multiple realizability. *Behavior and Philosophy* 23:29-39.

Block, N. 1997. Anti-reductionism slaps back. *Philosophical Perspectives* 11:107-32.

Bolender, J. 1995. Is multiple realizability compatible with antireductionism? *Southern Journal of Philosophy* 33:129-42.

Endicott, R. P. 1991. Macdonald on type reduction via disjunction. *Southern Journal of Philosophy* 29:209-14.

Endicott, R. P. 1989. On physical multiple realization. *Pacific Philosophical Quarterly* 70:212-24.

Endicott, R. P. 1993. Species-specific properties and more narrow reductive strategies. *Erkenntnis* 38:303-21.

On species-specific reductions. These can't reduce standard psychological properties, and problems with intra-species multiple realization can't be circumvented without giving up property reduction for token event identity.

Endicott, R. P. 1998. Collapse of the new wave. *Journal of Philosophy* 95:53-72.

Fodor, J. 1997. Special sciences: Still autonomous after all these years. *Philosophical Perspectives* 11:149-63.

Francescotti, R. M. 1997. What multiple realizability does not show. *Journal of Mind and Behavior* 18:13-28.

The anti-reductionist argument assumes that functional properties aren't physical properties (not even extrinsic physical properties). This, not multiple realizability, does the work.

Heil, J. 1999. Multiple realizability. *American Philosophical Quarterly* 36:189-208.

Kim, J. 1992. Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research* 52:1-26. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Multiple realization is compatible with reductionism. Jade (= jadeite or nephrite) isn't a scientific kind, and neither are multiply realizable mental properties. So there's no global psychology, just lots of local reductions.

Kistler, M. 1999. Multiple realization, reduction and mental properties. *International Studies in the Philosophy of Science* 13.

Nelson, A. 1985. Physical properties. *Pacific Philosophical Quarterly* 66:268-82.

Some comments on Wilson 1985: some special-science properties may be relevantly different in kind from his expanded physical properties.

Macdonald, C. 1992. Psychological type-type reduction via disjunction. *Southern Journal of Philosophy* 30:65-69.

Mucciolo, L. 1974. The identity thesis and neuropsychology. *Nous* 8:327-42.

Argues contra Fodor and Block that neurological equipotentiality doesn't refute type materialism. Mental states may not be anatomically defined neural states, but they may be more abstract neural holograms.

Nasrin, M. 2000. Multiple realizability: Also a difficulty for functionalism. *Journal of Consciousness Studies* 7:25-34.

Ross, P. A. 1999. The limits of physicalism. *Philosophy of Science* 66:94-116. Schwartz, J. 1992. Who's afraid of multiple realizability?: Functionalism, reductionism, and connectionism. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

Shagrir, O. 1998. Multiple realization, computation and the taxonomy of psychological states. *Synthese* 114:445-461.

Shapiro, L. 2000. Multiple realizations. *The Journal of Philosophy* 97:635-654.

Wilson, M. 1985. What is this thing called `pain'? -- The philosophy of science behind the contemporary debate. *Pacific Philosophical Quarterly* 66:227-67.

Argues for type-type identities and for an expanded view of the physical, as properties from physics exhibit the same sort of multiple realizability as functional properties. Sophisticated, with many interesting examples.

Zangwill, N. 1995. Supervenience, reduction, and infinite disjunction. *Philosophia* 24:321-30.

3.2b

Nonreductive Materialism [see also 3.5d]

Barrett, J. 1995. Causal relevance and nonreductive physicalism. *Erkenntnis* 42:339-62.

Beckermann, A. 1992. Reductive and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

On varieties of physicalism with respect to reduction: semantic physicalism, identity theory, supervenience, and the denial of emergence. Advocates a version on which physical states realize mental states.

Beckermann, A, Flohr, H. & Kim, J. (eds) 1992. *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Boyd, R. 1980. Materialism without reductionism: What physicalism does not entail. In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol 1. MIT Press.

Dupre, J. 1988. Materialism, physicalism, and scientism. *Philosophical Topics* 16:31-56.

Arguing for a pluralistic conception. With criticism of Churchland's reductionism, Davidson's token identity, and more generally reverential "scientism". Reductionist explanation is not the general rule.

Ellis, R. 2000. Consciousness, self-organization, and the process-substratum relation: Rethinking nonreductive physicalism. *Philosophical Psychology* 13:173-190.

Fodor, J. A. 1974. Special sciences. *Synthese* 28:97-115. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Psychological kinds can't be reduced to physical kinds, due to cross-classification, although token physicalism still holds. How to maintain the generality of physics without a reductionist unity of science.

Francescotti, R. M. 1998. The nonreductionist's troubles with supervenience. *Philosophical Studies* 89:105-24.

Horgan, T. 1993. Nonreductive materialism and the explanatory autonomy of psychology. In (S. Wagner & R. Warner, eds) *Naturalism: A Critical Appraisal*. University of Notre Dame Press.

Gives four constraints on interlevel connections, and some arguments against reductionism and for the autonomy of psychology. Argues that supervenience facts are themselves in need of explanation.

Horgan, T. 1994. Nonreductive materialism. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Kernohan, A. 1988. Non-reductive materialism and the spectrum of mind-body identity theories. *Dialogue* 27:475-88.

Classifying psychophysical theories by the status (necessary, lawful, anomalous, false) of psychophysical/psychological generalizations. Defending autonomous monism: nonreductive materialism with psychological laws.

Kim, J. 1989. The myth of non-reductive materialism. *Proceedings and Addresses of the American Philosophical Association* 63(3):31-47. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Somewhat loose arguments that non-reductive physicalist realism is untenable. Anomalous monism makes the mental irrelevant, functionalism is compatible with species-specific reduction, and supervenience is weak or reductive.

Kim, J. 1992. The nonreductivist's trouble with mental causation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Argues that nonreductive materialism implies downward causation (as the mental has more causal powers than the physical alone), and that downward causation violates the causal closure of the physical.

Kim, J. 1992. "Downward causation" in emergentism and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Argues that nonreductive materialism is just like 1930s emergentism, with the mental contributing new causal powers, and so implies downward causation.

Kirk, R. 1996. How physicalists can avoid reductionism. *Synthese* 108:157-70.

Contra Kim, physicalists can avoid reduction by embracing strict implication.

Margolis, J. 1978. *Persons and Minds: The Prospects of Non-Reductive Materialism*. D. Reidel.

- Marras, A. 1993. Psychophysical supervenience and nonreductive materialism. *Synthese* 95:275-304.
- Marras, A. 1994. Nonreductive materialism and mental causation. *Canadian Journal of Philosophy* 24:465-93.
- Melnyk, A. 1995. Two cheers for reductionism, or, the dim prospects for nonreductive materialism. *Philosophy of Science* 62:370-88.
- Melnyk, A. 1998. The prospects for Kirk's nonreductive physicalism. *Australasian Journal of Philosophy* 76:323-32.
- Loar, B. 1992. Elimination versus nonreductive physicalism. In (D. Charles & K. Lennon, eds) *Reduction, Explanation and Realism*. Oxford University Press.
- Papineau, D. 1992. Irreducibility and teleology. In (D. Charles & K. Lennon, eds) *Reduction, Explanation and Realism*. Oxford University Press.

Non-reductive physicalism is a mystery unless we invoke teleology.

- Pereboom, D. & Kornblith, H. 1991. The metaphysics of irreducibility. *Philosophical Studies* 63:125-45.

Explicating anti-reductionism: mental causal powers are constituted of physical causal powers, but aren't type- or token-identical to them. Against arguments from local reduction, neuroscience, explanatory exclusion, etc.

- Silvers, S. 1997. Nonreductive naturalism. *Theoria* 12:163-84.

- Smith, A. D. 1993. Non-reductive physicalism? In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

A careful discussion of how to characterize physicalism, in terms of identity or supervenience, and argues that physicalism must reduce (bowdlerize) qualia to something they are not, as physicalism requires topic-neutral analyses.

- Ten Elshof, G. 1997. Supervenient difficulties with nonreductive physicalism: A critical analysis of supervenience physicalism. *Kinesis* 24:3-22.

- van Gulick, R. 1992. Nonreductive materialism and the nature of intertheoretical constraint. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

On how a nonreductive materialism can handle problems about mental causation, psychophysical dependencies, and qualia. A teleofunctionalist view with different conceptual frameworks, but mental

properties physically realized.

Wedgwood, R. 2000. The price of non-reductive physicalism. *Nous* 34:400-421.

3.2c Reduction in Psychology

Bickle, J. 1995. Psychoneural reduction of the genuinely cognitive: Some accomplished facts. *Philosophical Psychology* 8:265-85.

Argues that cognitive theories have already been reduced to neurobiology in some domains, such as associative learning.

Churchland, P. M. 1982. Is 'thinker' a natural kind? *Dialogue* 21:223-38.

Psychology shouldn't be autonomous from natural science. By analogy with biology, nature provides (a) conceptual insight, and (b) real constraints, e.g. thermodynamic ones. Biology and psychology are continuous.

Gaito, J. 1960. Description, explanation, and reductionism in psychology. *Psychological Reports* 6:203-5.

Gaito, J. & Leonard, D. 1965. Philosophical and empirical reductionism in psychology. *Journal of General Psychology* 72:69-75.

Hardcastle, V. G. 1992. Reduction, explanatory extension, and the mind/brain sciences. *Philosophy of Science* 59:408-28.

The relationship between psychology and neuroscience is best characterized not by reduction but by explanatory extension, where each field is enriched by the other. With a number of examples from recent empirical work.

Hyland, M. E. 1995. Against nomological reductionism in psychology: A response to Robinson. *New Ideas in Psychology* 13:9-11.

Jessor, R. 1958. The problem of reductionism in psychology. *Psychological Review* 65:170-78.

Marras, A. 1990. Reduction in psychology. *Acta Analytica* 6:65-78.

Martindale, R. L. & Seidel, R. J. 1959. Reductionism: Its prodigal encores. *Psychological Reports* 5:213-16.

Montgomery, R. 1990. The reductionist ideal in cognitive psychology. *Synthese* 85:279-314.

Anti-reductionism needn't be ad hoc (contra Churchland). Although evolution provides some pressure for 1-1 psychophysical mappings, there are significant countervailing forces, e.g. in vision, memory, learning, and language use.

Olszewsky, T. M. 1975. Dispositions and reductionism in psychology. *Journal for the Theory of Social Behavior* 5:129-44.

Putnam, H. 1974. Reductionism and the nature of psychology. *Cognition* 2:131-46.

Richardson, R. C. 1999. Cognitive science and neuroscience: New wave reductionism. *Philosophical Psychology* 12:297-307.

Sloane, E. H. 1945. Reductionism. *Psychological Review* 52:214-23.

3.2d

Reduction, Misc

Beckermann, A. 1997. Property physicalism, reduction, and realization. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Bickle, J. 1996. New wave psychophysical reductionism and the methodological caveats. *Philosophy and Phenomenological Research* 56:57-78.

Bickle, J. 1997. *Psychoneural Reductionism: The New Wave*. MIT Press.

Brooks, D. H. M. 1994. How to perform a reduction. *Philosophy and Phenomenological Research* 54:803-14.

Reduction comes to supervenience plus explicability. Thus biconditionals, multiple realizability, etc, are irrelevant. Biology is already reduced (mostly via functional explanation), and psychology looks promising. Nice.

Bunzl, M. 1987. Reductionism and the mental. *American Philosophical Quarterly* 24:181-9.

On the links between supervenience, reduction, and explanation. Supervenience is compatible with reductive explanation of a localized variety. We don't need laws, but explanatory links.

Causey, R. L. 1972. Attribute identities in microreductions. *Journal of Philosophy* 69:407-22.

Combes, R. 1988. Ockhamite reductionism. *International Philosophical Quarterly* 28:325-36.

Foss, J. 1995. Materialism, reduction, replacement, and the place of consciousness in science. *Journal of Philosophy* 92:401-29.

Hill, C. S. 1984. In defense of type materialism. *Synthese* 59:295-320.

Kitcher, P. S. 1980. How to reduce a functional psychology. *Philosophy of Science* 47:134-40.

Contra Richardson 1979, a purely functional psychology is irreducible. The genetics analogy is misleading; multiple realizations can't explain high-level laws.

Papineau, D. 1985. Social facts and psychological facts. In (G. Currie & A. Musgrave, eds) *Popper and the Human Sciences*. Martinus Nijhoff.

Mind is not reducible to body, but societies reduce to individuals. Multiple realization is in tension with predictability. Natural selection resolves the tension for the mental, but cannot for the social.

Richardson, R. C. 1979. Functionalism and reductionism. *Philosophy of Science* 46:533-58.

Argues that functionalism is compatible with reductionism, by analogies. Genetics has multiple realization and multiple function; reduction doesn't require biconditionals. With remarks on the de facto autonomy of psychology.

Richardson, R. C. 1982. How not to reduce a functional psychology. *Philosophy of Science* 49:125-37.

Response to Kitcher 1980. Reductions are usually domain-specific, and high-level regularities are indeed explained.

Sarkar, S. 1992. Models of reduction and categories of reductionism. *Synthese* 91:167-94.

Wimsatt, W. 1976. Reductionism, levels of organization, and the mind-body problem. In (G. Globus, ed) *Consciousness and the Brain*. Plenum Press.

Excellent coverage of the notion of level and its applicability to mind.

3.3

Other Psychophysical Relations

3.3a

Physicalism [see also 1.3, 1.7b, 3.1b, 3.2b, 3.5]

Crane, T. 1991. All God has to do. *Analysis* 51:235-44.

If there are no contingent psychophysical laws, then there are no mental properties. So physicalism/supervenience is false; God had extra work to do.

Crane, T. 1993. A definition of physicalism: Reply to Pettit. *Analysis* 53:224-27.

Crane, T. & Mellor, D. H. 1990. There is no question of physicalism. *Mind* 99:185-206.

Physical sciences have no ontological authority over the mental. Considers and dismisses arguments from laws, causation, reduction, supervenience.

Daly, C. 1995. Does physicalism need fixing? *Analysis* 55:135-41.

Francescotti, R. 2000. Ontological physicalism and property pluralism: Why they are incompatible. *Pacific Philosophical Quarterly* 81:349-362.

Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*. Holder-Pichler-Tempsky.

On why materialism requires conceptual analysis to locate mental properties in the natural world. Even a posteriori necessary connections have to be backed by a priori links. With remarks on supervenience. A nice paper.

Kirk, R. 1979. From physical explicability to full-blooded materialism. *Philosophical Quarterly* 29:229-37.

If every physical events has a physical explanation, and the mental is causally efficacious, then mental facts are strictly implied by physical facts. A nice argument.

Kirk, R. 1982. Physicalism, identity, and strict implication. *Ratio* 24:131-41.

Materialism doesn't need a identity thesis. The requirement that mental facts are entailed by physical facts plays the role played by Kripke's requirement of necessary identity, and is more reasonable.

Kirk, R. 1996. Physicalism lives. *Ratio* 9:85-89.

Nothing in the arguments of Crane and Mellor 1990 count against a physicalism based on strict implication.

Madell, G. 1988. *Mind and Materialism*. Edinburgh University Press.

On the problems posed for materialism by intentionality, autonomy, awareness, and indexicality.

Tentatively advocates a Cartesian position.

McGinn, C. 1980. Philosophical materialism. *Synthese* 44:173-206. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Melnyk, A. 1994. Being a physicalist: How and (more importantly) why.

Advocates "realization physicalism": all properties are either physical or functional properties realized by physical ones. This achieves unity between sciences better than alternatives, and avoids overdetermination.

Melnyk, A. 1996. Formulating physicalism: Two suggestions. *Synthese* 105:381-407.

Discusses two formulations of physicalism: requiring high-level properties to be disjunctions of physical states, or to be functional properties realized physically. Tentatively endorses the latter.

Melnyk, A. 1997. How to keep the 'physical' in physicalism. *Journal of Philosophy* 94:622-637.

Montero, B. 1999. The body problem. *Nous* 33:183-200.

Moser, P. K. 1996. Physicalism and mental causes: Contra Papineau. *Analysis* 56:263-67.

Nagel, E. 1949. Are naturalists materialists? *Journal of Philosophy* 42:515-53.

Papineau, D. 1994. *Philosophical Naturalism*. Blackwell.

Pettit, P. 1993. A definition of physicalism. *Analysis* 53:213-23.

Physicalism is the claim that (1) There are microphysical entities, (2) Microphysical entities constitute everything, (3) There are microphysical regularities, (4) Microphysical regularities govern everything.

Pettit, P. 1994. Microphysicalism without contingent micro-macro laws. *Analysis* 54:253-57.

Pettit, P. 1995. Microphysicalism, dottism, and reduction. *Analysis* 55:141-46.

Poland, J. 1994. *Physicalism: The Empirical Foundations*. Oxford University Press.

Ravenscroft, I. 1997. Physical properties. *Southern Journal Of Philosophy* 35:419-431.

Robinson, D. 1991. On Crane and Mellor's argument against physicalism. *Mind* 100:135-36.

Robinson, H. (ed) 1993. *Objections to Physicalism*. Oxford University Press.

Sheldon, W. H. 1946. Are naturalists materialists? *Journal of Philosophy* 43:197-209.

Sober, E. 1999. Physicalism from a probabilistic point of view. *Philosophical Studies* 95:135-74.

Sturgeon, S. 1998. Physicalism and overdetermination. *Mind* 107:411-432.

Wilkes, K. V. 1973. *Physicalism*. Routledge and Kegan Paul.

3.3b Token Identity [see also 3.5b, 3.5d]

Foster, J. 1994. The token-identity thesis. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Horgan, T. & Tye, M. 1985. Against the token identity theory. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

We individuate mental events by their causal role, but we can't individuate causes uniquely. So each mental event has multiple physical correlates, and token identity doesn't hold.

Hornsby, J. 1981. Which physical events are mental events? *Proceedings of the Aristotelian Society* 55:73-92.

Haugeland, J. 1982. Weak supervenience. *American Philosophical Quarterly* 19:93-103.

Supervenience doesn't imply token identity, and Davidson's argument for token identity equivocates on "event". But weak supervenience (mentally discernible worlds are physically discernible) is all we need. With nice examples.

Leder, D. 1985. Troubles with token identity. *Philosophical Studies* 47:79-94.

Physical/psychological token identity is no good: you can't individuate physical events without psychological predicates.

Lurie, Y. 1978. Correlating brain states with psychological phenomena. *Australasian Journal of Philosophy* 56:135-44.

Can't isolate the physical token of a belief, say, as it's always accompanied by other beliefs. Meaning doesn't come in discrete tokens.

Peacocke, C. 1979. Argument for token identity. In *Holistic Explanation*. Oxford University Press.

3.3c Emergence

Alexander, S. 1920. *Space, Time, and Deity*. Macmillan.

Atkin A. 1992. On consciousness: What is the role of emergence? *Medical Hypotheses* 38:311-14.

Beckermann, A. 1992. Supervenience, emergence, and reduction. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

On varieties of supervenience and of emergence, and of what is required for reduction. Argues that reduction involves general explanatory connections, whereas emergence involves unique and ultimate bridge laws.

Beckermann, A, Flohr, H. & Kim, J. (eds) 1992. *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Bedau, M. 1997. Weak emergence. *Philosophical Perspectives* 11:375-399.

Berenda, C. W. 1953. On emergence and prediction. *Journal of Philosophy* 50:269-74.

Bergmann, G. 1944. Holism, historicism, and emergence. *Philosophy of Science* 11:209-21.

Broad, C. D. 1925. *The Mind and its Place in Nature*. Routledge and Kegan Paul.

Bruntrup, G. 1998. Is psychophysical emergentism committed to dualism? The causal efficacy of emergent mental properties. *Erkenntnis* 48:133-51.

Bunge, M. 1977. Emergence and the mind. *Neuroscience* 2:501-9.

Garnett, A. C. 1942. Scientific method and the concept of emergence. *Journal of Philosophy* 39:477-86.

Haldane, J. 1996. The mystery of emergence. *Proceedings of the Aristotelian Society* 96:261-67.

A defence of radical emergence against Spencer-Smith 1995.

Hasker, W. 1982. Emergentism. *Religious Studies* 18:473-88.

Hasker, W. 1999. *The Emergent Self*. Cornell University Press.

Henle, P. 1942. The status of emergence. *Journal of Philosophy* 39:486-93.

Humphreys, P. 1996. Aspects of emergence. *Philosophical Topics* 24:53-71.

- Humphreys, P. 1997. How properties emerge. *Philosophy of Science* 64:1-17.
- Humphreys, P. 1997. Emergence, not supervenience. *Philosophy of Science Supplement* 64:337-45.
- Jones, D. H. 1972. Emergent properties, persons, and the mind-body problem. *Southern Journal of Philosophy* 10:423-33.
- Kekes, J. 1966. Physicalism, the identity theory, and the concept of emergence. *Philosophy of Science* 33:360-75.
- Kim, J. 1999. Making sense of emergence. *Philosophical Studies* 95:3-36.
- Klee, J. 1984. Microdeterminism and concepts of emergence. *Philosophy of Science* 51:44-63.
- Lovejoy, A. O. 1927. The meanings of "emergence" and its modes. In (E. S. Brightman, ed) *Proceedings of the Sixth International Congress of Philosophy*. Longmans, Green, and Co.
- Lowe, E. J. 2000. Causal closure principles and emergentism. *Philosophy* 75:571-586.
- Lowry, A. 1974. A note on emergence. *Mind* 83:276-77.
- Mackenzie, W. L. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:56-68.
- Margolis, J. 1986. Emergence. *Philosophical Forum* 17:271-95.
- McLaughlin, B. P. 1992. The rise and fall of British emergentism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.
- A careful account of British emergentism. Explicates their view of emergent causal powers and laws in terms of fundamental configurational forces, a coherent idea that turned out to be false. An excellent paper.
- Meehl, P. E. & Sellars, W. 1956. The concept of emergence. In (H. Feigl & M. Scriven, eds) *Minnesota Studies in the Philosophy of Science*, vol. 1. University of Minnesota Press.
- Morgan, C. L. 1923. *Emergent Evolution*. Williams and Norgate.
- Morris, C. R. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:49-55.
- Newman, D. 1996. Emergence and strange attractors. *Philosophy of Science* 63:245-61.
- Newman, D. V. 2001. Chaos, emergence, and the mind-body problem. *Australasian Journal of Philosophy* 79:180-96.

O'Connor, T. 1994. Emergent properties. *American Philosophical Quarterly* 31:91-104.

Argues against Alexander's and van Cleve's accounts of emergence, instead suggesting an account in terms of supervenience, non-structurality, and downward causation.

Pap, A. 1951. The concept of absolute emergence. *British Journal for the Philosophy of Science* 2:302-11.

Pepper, S. C. 1926. Emergence. *Journal of Philosophy* 23:241-45.

Peters, S. L. 1995. *Emergent Materialism: A Proposed Solution to the Mind-Body Problem*. University Press of America.

Pihlstrom, S. 1999. What shall we do with emergence? A survey of a fundamental issue in the metaphysics and epistemology of science. *South African Journal of Philosophy* 18:192-210.

Pluhar, E. 1978. Emergence and reduction. *Studies in History and Philosophy of Science* 9:279-89.

Ripley, C. 1984. Sperry's concept of consciousness. *Inquiry* 27:399-423.

An in-depth analysis of Sperry's views on consciousness. Sperry is not a dualist; he believes in "structural causation" based on emergent properties.

Rohrlich, F. 1997. Cognitive emergence. *Philosophy of Science Supplement* 64:346-58.

Rueger, A. 2000. Robust supervenience and emergence. *Philosophy of Science* 67:466-491.

Russell, E. S. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:39-48.

Schroder, J. 1998. Emergence: Non-deducibility or downwards causation? *Philosophical Quarterly* 48:433-52.

Silberstein, M. 1998. Emergence and the mind-body problem. *Journal of Consciousness Studies* 5:464-82.

Silberstein, M. & McGeever, J. 1999. The search for ontological emergence. *Philosophical Quarterly* 49:182-200.

Smart, J. J. C. 1981. Physicalism and emergence. *Neuroscience* 6:109-13.

Spencer-Smith, R. 1995. Reductionism and emergent properties. *Proceedings of the Aristotelian Society* 95:113-29.

Distinguishes radical, epistemic, and interactional emergence, favoring the latter. With consideration of qualia as a radical emergent.

Sperry, R. W. 1969. A modified concept of consciousness. *Psychological Review* 76:532-36.

Consciousness is an emergent property of brain dynamics that itself governs low-level flow of excitation. Midway between mentalism and materialism.

Sperry, R. W. 1991. In defense of mentalism and emergent interaction. *Journal of Mind and Behavior* 12:221-245.

Stace, W. T. 1939. Novelty, indeterminism, and emergence. *Philosophical Review* 48:296-310.

Stephan, A. 1992. Emergence -- a systematic look at its historical facets. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

On different ways of understanding emergence: as nonadditivity, novelty, nonpredictability, nondeducibility; and on problems about qualia and downward causation.

Stephan, A. 1997. Armchair arguments against emergence. *Erkenntnis* 46:305-14.

Teller, P. 1992. A contemporary look at emergence. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

An attempt to explicate "emergent" properties in terms of relational properties. Argues that even problem cases, e.g. space-time separation and phenomenal properties, might be treated this way.

van Cleve, J. 1990. Mind -- dust or magic? Panpsychism versus emergence. *Philosophical Perspectives* 4:215-226.

On Nagel 1979: emergence is more plausible than panpsychism. A construal of emergence as nomological supervenience without logical supervenience.

vandervert, L. R. 1991. On the modeling of emergent interaction: Which will it be, the laws of thermodynamics or Sperry's "wheel" in the subcircuitry? *Journal of Mind and Behavior* 12:535-39.

Wimsatt, W. C. 1997. Aggregativity: Reductive heuristics for finding emergence. *Philosophy of Science* 64:372-84.

Wynn, M. 1999. Emergent phenomena and theistic explanation. *International Philosophical Quarterly* 39:141-55.

3.3d

Dualism [see also 1.3f, 1.4e, 1.4f]

Averill, E. W. & Keating, B. 1981. Does interactionism violate a law of classical physics? *Mind* 90:102-7.

Interactionism is compatible with conservation of energy and momentum: the mind exerts a non-physical force on the brain.

Bricke, J. 1975. Interaction and physiology. *Mind* 84:255-9.

Efron, A. 1992. Residual asymmetric dualism: A theory of mind-body relations. *Journal of Mind and Behavior* 13:113-36.

Evans, S. 1981. Separable souls: A defense of minimal dualism. *Southern Journal of Philosophy* 19.

Herbert, R. T. 1998. Dualism/materialism. *Philosophical Quarterly* 48:159-75.

Larmer, R. 1986. Mind-body interactionism and the conservation of energy. *International Philosophical Quarterly* 26:277-85.

Various arguments about interactionism based on conservation of energy. C of E only applies to causally isolated systems, so objections beg the question.

Lowe, E. J. 1992. The problem of psychophysical causation. *Australasian Journal of Philosophy* 70:263-76.

Argues that there can be interaction without breaking physical laws: e.g. by basic psychic forces, or by varying physical constants, or especially by arranging fractal trees of physical causation leading to behavior.

Lowe, E. J. 1993. The causal autonomy of the mental. *Mind* 102:629-44.

Mills, E. 1996. Interactionism and overdetermination. *American Philosophical Quarterly* 33:105-115.

Argues that interactionist dualism is compatible with the causal closure of the physical, if we allow causal overdetermination; and there is a strong case for the latter.

Mills, E. 1997. Interactionism and physicality. *Ratio* 10:169-83.

O'Leary-Hawthorne, J. & McDonough, J. K. 1998. Numbers, minds, and bodies: A fresh look at mind-body dualism. *Philosophical Perspectives* 12:349-371.

- Pap, A. 1952. Semantic analysis and psychophysical dualism. *Mind*.
- Pietroski, P. M. 1994. Mental causation for dualists. *Mind and Language* 9:336-66.
- Popper, K. R. 1953. Language and the body-mind problem: A restatement of interactionism. In *Proceedings of the 11th International Congress of Philosophy. Reprinted in Conjectures and the Growth of Scientific Knowledge. Basic Books, 1962.*
- Popper, K. R. 1955. A note on the body-mind problem. *Analysis* 15:131-35.
- Popper, K. R. 1977. Natural selection and the emergence of mind.
- Scheffler, I. 1950. The new dualism: Psychological and physical terms. *Journal of Philosophy*.
- Sellars, W. 1954. A note on Popper's argument for dualism. *Analysis* 15:23-24.
- Sussman, A. 1981. Reflection on the chances for a scientific dualism. *Journal of Philosophy* 78:95-118.

Dualism is an empty hypothesis. Everything must be matter, though we may have to expand the notion of matter.

Richardson, R. C. 1982. The 'scandal' of Cartesian dualism. *Mind* 91:20-37.

van Rooijen, K. 1987. Interactionism and evolution: A critique of Popper. *British Journal for the Philosophy of Science* 38:87-92.

3.3e Psychophysical Relations, Misc

- Campbell, K. 1983. Abstract particulars and the philosophy of mind. *Australasian Journal of Philosophy* 61:129-41.
- Caston, V. 1997. Epiphenomenalisms ancient and modern. *Philosophical Review* 106:309-363.
- Hedman, C. G. 1970. On correlating brain states with psychological states. *Australasian Journal of Philosophy* 48:247-51.
- Heil, J. 1992. *The Nature of True Minds*. Cambridge University Press.
- Honderich, T. 1981. Psychophysical law-like connections and their problems. *Inquiry* 24:277-303.

Defending lawlike connections between physical states & conscious occurrents. Contra anomalous monism

and identity theory for occurrents. But occurrents may not be causally efficacious. Comments by Wilson/Sprigge/Mackie/Stich.

McGinn, C. 1978. Mental states, natural kinds and psychophysical laws. *Proceedings of the Aristotelian Society* 52:195-220. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Argues that mental kinds are not natural kinds, and don't have real essences but nominal essences. For this reason, there are no psychophysical laws. With remarks on psychological laws, and the role of behavior.

Schectman, M. 1997. The brain/body problem. *Philosophical Psychology* 10:149-64.

Scheerer, E. 1994. Psychoneural isomorphism: Historical background and current relevance. *Philosophical Psychology* 7:183-210.

Skillen, A. 1984. Mind and matter: a problem which refuses dissolution. *Mind* 93:514-26.

Physical completeness, mental causation, non-reductionism are inconsistent. Ryle and Putnam are closet dualists, and Davidson's an epiphenomenalist.

Steward, H. 1997. *The Ontology of Mind: Events, Processes, and States*. Oxford University Press.

Tye, M. 1989. *The Metaphysics of Mind*. Cambridge University Press.

van Gelder, T. 1998. Monism, dualism, pluralism. *Mind and Language* 13:76-97.

3.4

Functionalism [see also 1.8, 4.6]

3.4a Causal Role Functionalism (Armstrong/Lewis)

Armstrong, D. M. 1968. *A Materialist Theory of the Mind*. Routledge and Kegan Paul.

Mental states should be analyzed as states that are apt to bring about certain kinds of behavior. Analysis of all kinds of mental states as such. With comments on dualism, behaviorism, identity theory, and consciousness.

Armstrong, D. M. 1970. The nature of mind. In (C. Borst, ed) *The Mind/Brain Identity Theory*. Macmillan. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Mental states are internal states that are apt to cause certain behaviors. A synthesis between the "thesis" of idealism and the "antithesis" of behaviorism. With defense against objections from consciousness.

Clark, A. 1986. Psychofunctionalism and chauvinism. *Philosophy of Science* 53:535-59.

Psychofunctionalism can evade chauvinism by specifying different functional identifications within each species. Applying same mental terms to each is justified by theory similarity; but it still isn't analytic functionalism.

Goldstein, I. 1994. Identifying mental states: A celebrated hypothesis refuted. *Australasian Journal of Philosophy* 72:46-62.

Against functionalism: experiences have intrinsic introspectible acausal properties, such as duration, felt location, and unpleasantness. Both analytic and empirical functionalism fail.

Horgan, T. 1984. Functionalism and token physicalism. *Synthese* 59:321-38.

Formalizing versions of functionalism, and seeing which entail token physicalism and/or type physicalism. On the most plausible versions, we have token physicalism without type physicalism.

Hornsby, J. 1984. On functionalism, and on Jackson, Pargetter, and Prior on functionalism. *Philosophical Studies* 46:75-96.

Jackson, F. , Pargetter, R. & Prior, E. W. 1982. Functionalism and type-type identity theories. *Philosophical Studies* 42:209-25.

Functionalism is compatible with type identity, as e.g. "pain" designates the state-type that fills the right functional role in an organism at a given time, i.e. a brain state. Contra Kripke, pain is not a rigid designator.

Kernohan, A. 1990. Lewis's functionalism and reductive materialism. *Philosophical Psychology* 3:235-46.

Argues that Lewis's functionalism founders on the specification of behavior. Described intentionally => non-materialist; physically => chauvinist.

Lewis, D. 1966. An argument for the identity theory. *Journal of Philosophy* 63:17-25. Reprinted in *Philosophical Papers, Vol. 1* (Oxford University Press, 1980).

Causal roles are definitive of mental states. Since physical states fill these causal roles (by the explanatory adequacy of physics), mental states are physical states.

Lewis, D. 1972. Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50:249-58. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Mental states can be defined, via a Ramsey-sentence analysis of the platitudes of folk psychology, as entities that fill causal roles specified by the analysis. These fillers turn out to be physical.

Lewis, D. 1978. Mad pain and martian pain. In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol. 1. MIT Press.

Accounting for both pains that don't play the usual causal role and for pains that are realized in different substances, by a mixed theory: pain is the physical state that typically occupies a certain causal role in a population.

McGinn, C. 1980. Functionalism and phenomenalism: A critical note. *Australasian Journal of Philosophy* 58:35-46. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Functionalism (reducing the mental to its effects on the physical) is no more plausible than phenomenalism (reducing the physical to its effects on the mental).

Owens, J. 1982. The failure of Lewis's functionalism. *Philosophical Quarterly* 36:159-73.

Lewis's original theory leads to Kripkean reference-fixing, so chauvinism. Token functionalism can't deal with paralytics. Species-relative functionalism fails as pain is intrinsic, not extrinsic.

Rogler, E. 2000. On David Lewis' philosophy of mind. *Protosociology* 14:285-311.

Sayward, C. 1995. Taking actions seriously. *Behavior and Philosophy* 23:51-60.

Shoemaker, S. 1981. Some varieties of functionalism. *Philosophical Topics* 12:93-119. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Fleshing out Ramsey-sentence functionalism; against Lewis's "mad pain" mixed theory; relating functionalism to the causal theory of properties. Empirical functionalism is chauvinistic so probably false. A terrific, in-depth paper.

Tye, M. 1983. Functionalism and type physicalism. *Philosophical Studies* 44:161-74.

Contra Lewis: Functionalism isn't compatible with type physicalism. There are intra-population difficulties with species-relative construals, and individual-relative construals can still have multiple fillers.

3.4b Machine Functionalism (Putnam) [see also 4.8]

Putnam, H. 1960. Minds and machines. In (S. Hook, ed) *Dimensions of Mind*. New York University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

The relationship between mental and physical states is just like that between logical and structural states of Turing Machines, so no great mystery. With comments on privacy and semantic analysis.

Putnam, H. 1967. The nature of mental states. In (Capitan & Merrill, eds) *Art, Mind, and Religion*.

Pittsburgh University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Why mental states are more likely to be functional states (in probabilistic automata) than brain states or behavioral dispositions.

Putnam, H. 1967. The mental life of some machines. In (H. Castaneda, ed) *Intentionality, Minds and Perception*. Wayne State University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

On explaining behavior via TM states, e.g. explaining preference via utility functions. Logical behaviorism assumes rational preference functions. Functional organization is what matters, not physical make-up.

Putnam, H. 1975. Philosophy and our mental life. In *Mind, Language, and Reality*. Cambridge University Press.

Psychological states aren't TM states after all: we have lots of psych states at once; they depend on learning/memory; disjunctions of TM states are no good. But functional organization rather than physics is still what counts.

Putnam, H. 1987. *Representation and Reality*. MIT Press.

Type functionalism isn't any better than type physicalism, as mental states can be multiply realized as functional states. With what in common?

Lycan, W. G. 1974. Mental states and Putnam's functionalist hypothesis. *Australasian Journal of Philosophy* 52:48-62.

On abstract vs. physical TMs: Putnam should say that mental states are physical TM states. But then functionalism is compatible with physicalism. On the relation between Putnam's and Armstrong's functionalism.

Lycan, W. G. 1979. A New Lilliputian argument against machine functionalism. *Philosophical Studies* 35:279-87.

If machine functionalism were true, a homunculus-head would have all the mental states of its homunculus (by the definition of "realization"), which is absurd.

Lycan, W. G. 1983. The moral of the New Lilliputian argument. *Philosophical Studies* 43:277-80.

Reply to Elugardo 1983: so how do you specify what count as inputs/outputs?

Elugardo, R. 1981. Machine functionalism and the New Lilliputian argument. *Pacific Philosophical Quarterly* 62:256-61.

Criticism of Lycan 1979, and a re-making of the argument.

Elugardo, R. 1983. Machine realization and the New Lilliputian argument. *Philosophical Studies* 43:267-75.

Lycan's New Lilliputian argument fails as inputs/outputs for the homunculus are not the same as inputs/outputs for the full system.

Kane, R. H. 1966. Turing machines and mental objects

Nelson, R. 1974. Mechanism, functionalism, and the identity theory. *Journal of Philosophy* 73:365-86.

Argues for mechanism rather than functionalism. Criticizes Putnam for hypostasizing mental states, which are disanalogous to mental states. Defending mechanism against Kalke's & Rorty's objections.

Rorty, R. 1972. Functionalism, machines and incorrigibility. *Journal of Philosophy* 69:203-20.

Logical states don't give us any understanding of mind over and above what the function/structure distinction gives us. In particular, it doesn't help with the understanding of privacy and incorrigibility.

Tomberlin, J. 1965. About the identity theory. *Australasian Journal of Philosophy* 43:295-99.

Contra Putnam: logical states are not physical states, and utterances about them are not about physical states.

Wagner, S. J. 1988. The liberal and the lycanthrope. *Pacific Philosophical Quarterly* 69:165-74.

Contra Lycan: machine functionalism can handle Bolivia and CRT cases by a causal/counterfactual account, and Lilliputian case by assigning mental states to minds, not bodies.

3.4c Functionalism, Miscellaneous

Adams, F. 1979. Properties, functionalism, and the identity theory. *Eidos* 1:153-79.

Bealer, G. 1978. An inconsistency in functionalism. *Synthese*.

A formal argument showing that functional definitions are equivalent to behavioral definitions.

Bealer, G. 1985. Mind and anti-mind: Why thinking has no functional definition. *Midwest Studies in Philosophy* 9:283-328.

Bechtel, W. 1984. Autonomous psychology: What it should and should not entail. *Philosophy of Science*

Association 1984, 1:43-55.

The functional level is the appropriate level for psychology, but neurophysiological facts constrain this level and are thus relevant.

Ben-Yami, H. 1999. An argument against functionalism. *Australasian Journal of Philosophy* 77:320-324.

Biro, J. I. & Shahan, R. W. (eds) 1982. *Mind, Brain and Function*. Oklahoma University Press.

Ten papers on functionalism. Originally was *Philosophical Topics*, volume 12.

Block, N. 1980. Functionalism. In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol. 1. MIT Press.

Distinguishes varieties of functionalism, e.g. machine and Ramsey-sentence functionalism; and compares to behaviorism. With a historical overview, and arguments for why functionalism is incompatible with physicalism.

Block, N. 1978. Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9:261-325. Reprinted in *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Distinguishes analytic and empirical functionalism. Both have problems with absent qualia, and inputs/outputs. Analytic functionalism has problems with paralytics, etc; empirical functionalism has problems with Martians.

Block, N. & Fodor, J. A. 1972. What psychological states are not. *Philosophical Review* 81:159-81. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Mental states are not physical or behavioral states; could they be functional states? With various arguments against type identity, and against machine-table functionalism.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72:741-64. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

On the role of functional explanation versus other kinds of explanation. Functionalism applies an analytic, not subsumptive strategy.

David, M. 1997. Kim's functionalism. *Philosophical Perspectives* 11:133-48.

Fischer, J. 1985. Functionalism and propositions. *Philosophical Studies* 48:295-311.

Fodor, J. A. 1968. Materialism. In *Psychological Explanation*. Random House.

On mental state as inferred theoretical entities, individuated according to their function (cf. valve-lifters). Psychology and neuroscience will mutually constrain each other, giving a relation more complex than reduction.

Gendron, B. 1970. On the relation of neurological and psychological theories: A critique of the hardware thesis. *Boston Studies in the Philosophy of Science* 8:483-95.

Argues that functional explanation are reducible to structural explanations.

Hornsby, J. 1986. Physicalist thinking and conceptions of behaviour. In (P. Pettit & J. McDowell, eds) *Subject, Thought, and Context*. Oxford University Press.

Hoy, R. C. 1980. Dispositions, logical states, and mental occurrents. *Synthese* 44:207-40.n

Kalke, W. 1969. What's wrong with Fodor's and Putnam's functionalism. *Nous* 3:83-93.

There's no absolute functional/structural distinction, as it depends on how you choose boundaries and levels of abstraction.

Lycan, W. G. 1981. Form, function and feel. *Journal of Philosophy* 78:24-50.

Pursue a multi-leveled homuncular functionalism, with mental states characterized as states of teleologically identified subsystems. Even the identity theorist is a functionalist at a low level.

Malcolm, N. 1980. 'Functionalism' in philosophical psychology. *Proceedings of the Aristotelian Society* 80:211-30.

Pereboom, D. 1991. Why a scientific realist cannot be a functionalist. *Synthese* 88:341-58.

Scientific realism requires dispositions of kinds be explained by intrinsic properties. Neural/functional properties won't work, because of reductionism and circularity. Use intrinsic psychological properties instead.

Richardson, R. C. 1979. Functionalism and reductionism. *Philosophy of Science* 46:533-58.

Argues that functionalism is compatible with reductionism, by analogies. Genetics has multiple realization and multiple function; reduction doesn't require biconditionals. With remarks on the de facto autonomy of psychology.

Schiffer, S. 1986. Functionalism and belief. In (M. Brand & R. Harnish, eds) *The Representation of Knowledge and Belief*. University of Arizona Press.

Against functionalism for beliefs. Both common-sense functionalism and psychofunctionalism have problems with finding the right functional theory, distinguishing beliefs, perceptual input conditions, Twin

Earth, etc.

Shope, R. K. 1973. Functional equivalence and the defense of materialism. *Philosophical Forum* 4:500-12.

Sober, E. 1990. Putting the function back into functionalism. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Need teleological functionalism, not Turing Machine functionalism.

Sober, E. 1985. Panglossian functionalism and the philosophy of mind. *Synthese* 64:165-93.

van Gulick, R. 1982. Functionalism as a theory of mind. *Philosophy Research Archives* 185-204.

The structure/function distinction is level-relative, so physiology might be relevant even under functionalism. Problems with automata, and with causal connections to nonintentionally characterized behavior.

van Gulick, R. 1980. Functionalism, information and content. *Nature and System* 2:139-62. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Ward, A. 1989. Philosophical functionalism. *Behaviorism* 17:155-8.

Weckert, J. 1990. Functionalism's impotence. *Philosophical Inquiry* 32-43.

Wilkes, K. V. 1981. Functionalism, psychology and the philosophy of mind. *Philosophical Topics* 12:147-67.

Functionalism may be appropriate for cognitive psychology but not for folk psychology, due to differing goals. Neuroscience will play an important role in developing functional theories.

Zangwill, N. 1992. Variable realization: not proven. *Philosophical Quarterly* 42:214-19.

Argues that the possibility of multiple realization has not been established, whether by arguments from imagination, concepts, or empirical facts.

3.5

Other Psychophysical Theories

3.5a

Logical Behaviorism (Ryle, etc)

Ryle, G. 1949. *The Concept of Mind*. Hutchinson and Co.

The ancestor of most contemporary philosophy of mind. Among other things, argues that the "ghost in the machine" view of mind is a category mistake, and presents dispositional analyses of many mental concepts.

Bestor, T. W. 1979. Gilbert Ryle and the adverbial theory of mind. *Personalist* 60:233-42.

Campbell, C. A. 1953. Ryle on the intellect. *Philosophical Quarterly* 3:115-38.

Carnap, R. 1959. Psychology in physical language. In (Ayer, ed) *Logical Positivism*. Free Press.

Carrier, L. 1973. Professor Shaffer's refutation of behaviourism. *Mind* 80:249-52.

Chisholm, R. 1955. A note on Carnap's meaning analysis. *Philosophical Studies*.

Chisholm, R. 1952. Intentionality and the theory of signs. *Philosophical Studies*.

Chisholm, R. 1958. Sentences about believing. *Minnesota Studies in the Philosophy of Science* 2.

Dalrymple, H. 1977. Some logical muddles in behaviorism. *Southwestern Philosophical Studies* 2:64-72.

Ewing, A. C. 1953. Professor Ryle's attack on dualism. *Proceedings of the Aristotelian Society* 53:47-78.

Farrell, B. 1950. Experience. *Mind* 59:170-98.

Finn, D. R. 1971. Putnam and logical behaviourism. *Mind* 80:432-36.

Flanagan, O. J. & McCreadie-Albright, T. 1974. Malcolm and the fallacy of behaviorism. *Philosophical Studies* 26:425-30.

Geach, P. 1957. *Mental Acts*. Routledge and Kegan Paul.

Goudge, T. A. 1982. Ryle's last thoughts on thinking. *Dialogue* 21:125-32.

Graham, G. 19xx. Spartans and behaviorists. *Behaviorism* x:xx.

Defends behaviorism as a scientific hypothesis, so that conceivability arguments aren't relevant, and advocates "penetrability" behaviorism which can appeal to internal physical states.

- Hamer, C. 1970. Why Ryle is not a behaviourist. *Philosophical Studies (Ireland)* 17:7-25.
- Hamlyn, D. W. 1953. Behaviour. *Philosophy* 28:132-45.
- Hanson, N. R. 1952. Professor Ryle's "mind". *Philosophical Quarterly* 2:246-48.
- Heidelberger, H. 1966. On characterizing the psychological. *Philosophy and Phenomenological Research*.
- Kitchener, R. F. 1977. Behavior and behaviorism. *Behaviorism* 5:11-68.
- Jacquette, D. 1985. Logical behaviorism and the simulation of mental episodes. *Journal of Mind and Behavior* 6:325-332.
- Mace, C. A. 1949. Some implications of analytical behaviourism. *Proceedings of the Aristotelian Society*.
- Malcolm, N. 1954. Wittgenstein's *Philosophical Investigations*. *Philosophical Review* 43:530-9.
- Mandelbaum, M. 1958. Professor Ryle and psychology. *Philosophical Review* 67:522-30.
- McLaughlin, B. & O'Leary-Hawthorne, J. 1995. Dennett's logical behaviorism. *Philosophical Topics* 22:189-258.
- Miller, D. S. 1911. Is consciousness "a type of behaviour"? *Journal of Philosophy* 8:322-27.
- Miller, D. S. 1951. "Descartes myth" and "Professor Ryle's fallacy". *Journal of Philosophy*.
- Nelson, R. 1969. Behaviorism is false. *Journal of Philosophy* 66:417-52.
- Nelson, R. 1975. Behaviorism, finite automata, and stimulus-response theory. *Theory and Decision* 6:249-67.
- Oosthuizen, D. C. S. 1970. Phenomenological psychology. *Mind* 79:487-501.
- Park, S. 1994. Reinterpreting Ryle: A nonbehaviorist analysis. *Journal of the History of Philosophy* 32:265-90.
- Place, U. T. 1993. A radical behaviorist methodology for the empirical investigation of private events. *Behavior and Philosophy* 20:25-35.
- Price, H. H. 1960. Some objections to behaviorism. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

- Putnam, H. 1963. Brains and behavior. In (R. Butler, ed) *Analytical Philosophy: Second Series*. Blackwell. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).
- Quine, W. V. 1975. Mind and verbal dispositions. In (Guttenplan, ed) *Mind and Language*. Oxford University Press.
- Quine, W. V. 1980. Sellars on behaviorism, language, and meaning. *Pacific Philosophical Quarterly* 61:26-30.
- Robinson, H. 1982. Behaviorism and stimulus materialism. In *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.
- Rowlands, M. 1991. A defense of behaviorism. *Behavior and Philosophy* 19:93-100.
- Ryle, G. 1979. *On Thinking*. Blackwell.
- Scriven, M. 1956. A study of radical behaviorism. *Minnesota Studies in the Philosophy of Science* 1:88-130.
- Sellars, W. 1952. Mind, meaning, and behavior. *Philosophical Studies*.
- Shuford, H. 1966. Logical behaviorism and intentionality. *Theoria* 32:246-51.
- Skinner, B. F. 1945. The operational analysis of psychological terms. *Psychological Review* 52:270-78.
- Smart, J. J. C. 1959. Ryle on mechanism and psychology. *Philosophical Quarterly* 9:349-55.
- Stemmer, N. 1993. Behavioral materialism, the success of folk psychology, and the first-person case. *Behavior and Philosophy* 20:1-14.
- Vendler, Z. 1981. Ryle's thoughts on thinking. *Midwest Studies of Philosophy* 6:335-43.
- Weitz, M. 1951. Professor Ryle's "logical behaviourism". *Journal of Philosophy* 48:297-300.
- Whitely, C. A. 1961. Behaviourism. *Mind* 70:164-74.
- Wisdom, J. 1950. The concept of mind. *Proceedings of the Aristotelian Society* 50:189-204.
- Wittgenstein, L. 1953. *Philosophical Investigations*.
- Wright, J. N. 1959. Mind and the concept of mind. *Aristotelian Society Supplement* 33:1-22.

Ziff, P. 1958. About behaviourism. *Analysis* 18:132-6.

3.5b Identity Theory (Smart, etc) [see also 1.3c, 1.4g, 3.3b]

Feigl, H. 1958. The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science* 2:370-497. Reprinted as *The 'Mental' and the 'Physical'*. University of Minnesota Press, 1967.

Place, U. T. 1956. Is consciousness a brain process? *British Journal of Psychology* 47:44-50. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

The idea that consciousness is a brain process is logically coherent. It's a scientific hypothesis, not a necessary truth. On the "is" of composition vs the "is" of definition, and the fallacy of the internal phenomenal field.

Smart, J. J. C. 1959. Sensations and brain processes. *Philosophical Review* 68:141-56.

Defending the thesis that sensations are contingently identical to brain processes against various objections. Topic-neutral analysis of sensation reports. Materialism beats epiphenomenalism on grounds of simplicity.

Abelson, R. 1970. A refutation of mind-body identity. *Philosophical Studies* 18:85-90.

The number of possible mental states is infinite (think of any number), whereas there are only finitely many brain states, so they're not identical.

Armstrong, D. M. 1968. The headless woman and the defense of materialism. *Analysis* 29:48-49.

Likens the anti-materialist position to the "headless woman" fallacy: "I'm not aware the mental states are physical", so "I'm aware that mental states are non-physical".

Armstrong, D. M. 1973. Epistemological foundations for a materialist theory of mind. *Philosophy of Science* 40:178-93.

A prima facie case for materialism based on grounds of rational consensus, arising especially from common-sense and scientific evidence. Mental states exist (common-sense) but should be analyzed causally (evidence from science).

Aune, B. 1966. Feigl on the mind-body problem. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Baier, K. 1962. Smart on sensations. *Australasian Journal of Philosophy* 40:57-68.

Mental states are necessarily private, and so cannot be physical states, which are public. We have epistemological authority about our mental states.

Beloff, J. 1965. The identity hypothesis: A critique. In (J. R. Smythies, ed) *Brain and Mind*. Routledge and Kegan Paul.

Blumenfeld, J-B. 1979. Phenomenal properties and the identity theory. *Australasian Journal of Philosophy* 63:485-93.

Argues that phenomenal properties aren't needed to identify sensations with brain-states, and nor are topic-neutral analyses.

Borst, C. V. (ed) 1970. *The Mind/Brain Identity Theory*. Macmillan.

An anthology of central articles on the identity theory.

Bradley, M. C. 1963. Sensations, brain-processes, and colours. *Australasian Journal of Philosophy* 41:385-93.

Brandt, R. 1960. Doubts about the identity theory. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

Brandt, R. & Kim, J. 1967. The logic of the identity theory. *Journal of Philosophy* 66:515-537.

Arguing for an event-identity construal of the identity theory. Comparing the identity theory to the weaker "principle of simultaneous isomorphism". The only reason to accept the identity theory is ontological simplicity.

Brodbeck, M. 1966. Mental and physical: Identity versus sameness. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Candlish, S. 1970. Mind, brain, and identity. *Mind* 79:502-18.

Carney, J. 1971. The compatibility of mind-body identity with dualism. *Mind*.

Argues that the identity theory is compatible with linguistic dualism, as the mental and the physical may differ in intensional properties only.

Clarke, J. 1971. Mental structure and the identity theory. *Mind* 80:521-30.

Coburn, R. 1963. Shaffer on the identity of mental states and brain processes. *Journal of Philosophy* 60:89-92.

Coder, D. 1973. The fundamental error of central-state materialism. *American Philosophical Quarterly* 10:289-98.

On problems with theories that leave the nature of mind open a priori: how can we even understand the possibilities?

Cornman, J. 1962. The identity of mind and body. *Journal of Philosophy* 59:486-92.

Coburn, R. C. 1963. Shaffer on the identity of mental states and brain processes. *Journal of Philosophy* 60:89.

Location of mental states by convention (Shaffer 1961) won't work, as it (a) makes mental states public, and (b) conflicts with connections to behavior.

Crittenden, C. 1971. Ontology and mind-body identity. *Philosophical Forum* 2:251-70.

de Boer, R. 1976. Cartesian categories in mind-body identity theories. *Philosophical Forum* 7:139-58.

Double, R. 1981. Central state materialism. *Philosophical Studies (Ireland)* 28:229-37.

Enc, B. 1983. In defense of the identity theory. *Journal of Philosophy* 80:279-98.

Feigl, H. 1971. Some crucial issues of mind-body monism. *Synthese*.

Garnett, A. C. 1965. Body and mind: the identity thesis. *Australasian Journal of Philosophy* 43:77-81.

Grunbaum, A. 1972. Abelson on Feigl's mind-body identity thesis. *Philosophical Studies* 23:119-21.

Gustafson, D. F. 1963. On the identity theory. *Analysis* 24:30-32.

Hanratty, G. 1972. The identity theory of Herbert Feigl. *Philosophical Studies* 20:113-23.

Harris, E. E. 1966. The neural identity thesis and the person. *International Philosophical Quarterly* 6:515-37.

Hedman, C. G. 1970. On correlating brain states with psychological states. *Australasian Journal of Philosophy* 48:247-51.

Heil, J. 1970. Sensations, experiences, and brain processes. *Philosophy* 45:221-6.

Hinton, J. M. 1967. Illusions and identity. *Analysis* 27:65-76.

Hockutt, M. 1967. In defense of materialism. *Philosophy and Phenomenological Research* 27:366-85.

Hoffman, R. 1967. Malcolm and Smart on brain-mind identity. *Philosophy* 42: 128-36.

Joske, W. 1960. Sensations and brain processes: A reply to Professor Smart. *Australasian Journal of Philosophy* 38:157-60.

On topic-neutral reports, after-images, and after-radishes. Such a report requires epistemic access to physical resemblance, which we don't have.

Kim, J. 1966. On the psycho-physical identity theory. *American Philosophical Quarterly* 3:227-35.

There's no empirical support for identity, over and above that for correlation; and unity of science gives no reason to accept identity. The only reason might be that of ontological simplicity.

Kim, J. 1972. Phenomenal properties, psychophysical laws and the identity theory. *Monist* 56:178-92.

Deal with phenomenal properties by allowing only mental events, and eliminating mental objects. Identity theories needn't suppose psychophysical laws. With defense against multiple realizability arguments.

Kitcher, P. S. 1982. Two versions of the identity theory. *Erkenntnis* 17:213-28.

Recasting the identity theory and functionalism, using Kripkean theories of reference, so mental states can refer to physiological or psychological states that we don't yet understand; and qualia problems are handled better.

Lewis, D. 1965. An argument for the identity theory. *Journal of Philosophy* 63:17-25. Reprinted in *Philosophical Papers, Vol. 1* (Oxford University Press, 1980).

Mental states are defined by their causal roles. So, by the completeness of physics, they must be physical states.

Locke, D. 1971. Must a materialist pretend he's anaesthetized? *Philosophical Quarterly* 49:217-31.

On how materialism, as opposed to a double aspect view, can handle mental features -- by moving them into the world via a realist theory of perception. Remarks on identification of states. After-images, etc, cause problems.

Lockwood, M. 1984. Einstein and the identity theory. *Analysis*.

Using the special theory of relativity to show that if mental events have a temporal location, then they must have a spatial location.

Lubow, N. 1978. Mind-body identity and irreducible properties. *Philosophy Research Archives* 4:1240.

Luce, D. R. 1966. Mind-body identity and psycho-physical correlation. *Philosophy of Science* 17:1-7.

Malcolm, N. 1964. Scientific materialism and the identity theory. *Dialogue* 3:115-25.

The identity theory is meaningless, if identity is analyzed as spatiotemporal coincidence, as thoughts don't have location. Thoughts also require context. Even if identity holds, explaining brain doesn't imply explaining mind.

Macdonald, C. 1989. *Mind-Body Identity Theories*. Routledge.

Malcolm, N. 1964. Scientific materialism and the identity theory. *Dialogue*.

Margolis, J. 1965. Brain processes and sensations. *Theoria* 31:133-38.

Meehl, P. 1966. The compleat autocerebroscopist: A thought-experiment on Professor Feigl's mind-body identity thesis. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Mucciolo, L. 1974. The identity theory and criteria for the mental. *Philosophy and Phenomenological Research* 35:167-80.

Munsat, S. 1969. Could sensations be processes? *Mind* 78:247-51.

Sensations and processes have different logical type, so it is a priori impossible that they should be identical.

Nagel, T. 1965. Physicalism. *Philosophical Review* 74:339-56, 1965.

Noren, S. J. 1970. Identity, materialism, and the problem of the danglers. *Metaphilosophy* 4:318-44.

Noren, S. J. 1970. Smart's materialism: The identity thesis and translation. *Australasian Journal of Philosophy* 48:54-66.

Norton, R. 1964. On the identity of identity theories. *Analysis* 25:14-16.

Pepper, S. 1975. A split in the identity theory. In (C. Cheng, ed) *Philosophical Aspects of the Mind-Body Problem*. Hawaii University Press.

Pitcher, G. 1960. Sensations and brain processes: A reply to Professor Smart. *Australasian Journal of Philosophy* 38:150-7.

Identity requires explanation to be accepted, but Smart doesn't provide this. But one can deny identity without claiming dualism -- e.g. a "duck-rabbit" theory of mind/brain. With remarks on the completeness of descriptions.

Place, U. T. 1960. Materialism as a scientific hypothesis. *Philosophical Review* 69:101-4.

Contra Smart 1959: Materialism is a scientific hypothesis, if we accept certain logical criteria for what a sensation is; otherwise it's just false.

Place, U. T. 1972. Sensations and processes: A reply to Munsat. *Mind*.

Place, U. T. 1988. Thirty years on -- Is consciousness still a brain process? *Australasian Journal of Philosophy* 66:208-19.

Comparing contemporary materialism to Place's 1956 variety. With remarks on whether the thesis is empirical or a priori, and on deciding the issue between materialism and epiphenomenalism.

Place, U. T. 1989. Low claim assertions. In (J. Heil, ed) *Cause, Mind, and Reality: Essays Honoring C. B. Martin*. Kluwer.

Discusses a paper of Martin's and the genesis of the identity theory, with a focus on 'public' and 'private logic' and topic-neutral descriptions.

Presley, C. P. (ed) 1967. *The Identity Theory of Mind*. University of Queensland Press.

Puccetti, R. 1978. The refutation of materialism. *Canadian Journal of Philosophy* 8:157-62.

The identity theory must be false, as pain centers in vitro will not be pains. With a reply by G. Pearce and a rejoinder.

Ripley, C. 1969. The identity theory and scientific hypotheses. *Dialogue* 2:308-10.

Robinson, H. 1982. The disappearance theory. In *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Rosenbaum, S. 1977. The property objection and the principles of identity. *Philosophical Studies* 32.

Routley, R. & MaCrae, V. 1966. On the identity of sensations and physiological occurrences. *American Philosophical Quarterly* 3.

Schlagel, R. H. 1977. The mind-body identity impasse. *American Philosophical Quarterly* 14:231-37.

Scriven, M. 1966. The limitations of the identity theory. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

On the identity theory as a linguistic proposal, compatible with dualism; epiphenomenalism and parallelism must be false, leaving interactionism.

Sellars, W. 1965. The identity approach to the mind-body problem. *Review of Metaphysics* 18:430-51.

Shaffer, J. 1961. Could mental states be brain processes? *Journal of Philosophy* 58:813-22.

Mental states don't have a location, and brain processes do; but we could stipulate a location for mental states. With remarks on possible relations between mental and physical features, states, and concepts.

Shaffer, J. 1963. Mental events and the brain. *Journal of Philosophy* 60:160-6.

We identify mental events by noticing mental features that must be nonphysical, but still might be empirically reducible. Against topic-neutral definitions, and with response to Coburn 1963 on location.

Simon, M. A. 1970. Materialism, mental language, and the mind-body identity. *Philosophy and Phenomenological Research* 30:514-32.

Smart, J. J. C. 1960. Sensations and brain processes: A rejoinder to Dr. Pitcher and Mr. Joske. *Australasian Journal of Philosophy* 38:252-54.

Smart, J. J. C. 1961. Further remarks on sensations and brain processes. *Philosophical Review*.

Reply to Stevenson 1960: There are no irreducible mental properties; they reduce to physical properties via topic-neutral definitions.

Smart, J. J. C. 1962. Brain processes and incorrigibility. *Australasian Journal of Philosophy* 40:68-70.

Reply to Baier 1962: epistemological authority is compatible with materialism. Mental state reports are not completely incorrigible, though.

Smart, J. J. C. 1963. Materialism. *Journal of Philosophy* 60:651-62.

Defending topic-neutral analyses of mental reports, and arguing against Wittgensteinian behaviorism via brain-in-vat examples. With remarks on the appeal of materialism and on compatibility with ordinary language.

Smart, J. J. C. 1965. The identity thesis: A reply to Professor Garrett. *Australasian Journal of Philosophy* 43:82-3.

Smart, J. J. C. 1972. Further thoughts on the identity theory. *Monist* 56:177-92.

On some problems for the identity theory arising from the intensionality of mental states and from the appeal to properties, and on how to modify the translation form of the theory without embracing the disappearance version.

Smythies, J. R. 1994. Requiem for the identity theory. *Inquiry* 37:311-29.

Sosa, E. 1965. Professor Malcolm on "Scientific materialism and the identity theory". *Dialogue* 4:422-23.

Stevenson, J. T. 1960. 'Sensations and brain processes': A reply to J. J. C. Smart. *Philosophical Review* 69:505-10.

Identity theory implies nomological danglers, due to the irreducibility of defining mental properties.

Stoutland, F. 1971. Ontological simplicity and the identity hypothesis. *Philosophy and Phenomenological Research*.

The identity thesis isn't ontologically simpler than dualism: we still need a dualism of properties, and explanatory danglers. Not much turns on the issue, except in teleological explanation.

Sosa, E. 1965. Professor Malcolm on 'Scientific materialism and the identity theory'. *Dialogue* 3:422-23.

Contra Malcolm 1965: explaining brain will explain mind, if the explanation is conjoined with the identity statement. With rejoinder from Malcolm.

Swartz, N. 1974. Can the theory of contingent identity between sensation-states and brain-states be made empirical? *Canadian Journal of Philosophy* 3:405-17.

Swinburne, R. 1993. Are mental events identical with brain events? *American Philosophical Quarterly* 19:173-181.

Property identity theses fail due to meaning differences, and event identity these fail due to a lack of entailment relations. Rebuts objections from weaker identity criteria and analogies with scientific identification.

Taylor, C. 1967. Mind-body identity, a side issue? *Philosophical Review* 76:201-13.

Teichmann, J. 1967. The contingent identity of minds and brains. *Mind* 76:404-15.

Thalberg, I. 1978. A novel approach to mind-brain identity. *Philosophy of Science* 3:255-72.

Suggests a theory in which neural states are components of, but not identical to, overall psychological

states. This can accommodate raw feels if necessary as a further component, but is mostly materialistic.

Thomson, J. J. 1969. The identity theory. In (S. Morgenbesser, P. Suppes, & M. White, eds) *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*. St. Martin's Press.

Tomberlin, J. E. 1965. About the identity theory. *Australasian Journal of Philosophy* 53:295-9.

Watkins, J. W. N. 1978. A basic difficulty in the mind-brain identity hypothesis. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Weismann, D. 1965. A note on the identity thesis. *Mind* 74:571-77.

Whitely, C. H. 1970. The mind-brain identity hypothesis. *Philosophical Quarterly* 20:193-99.

Wolfe, J. & Nathan, G. J. 1968. The identity theory as a scientific hypothesis. *Dialogue* 7:469-72.

Ziedins, R. 1971. Identification of characteristics of mental events with characteristics of brain events. *American Philosophical Quarterly* 8:13-23.

3.5c Eliminative Materialism (Rorty, Feyerabend) [see also 1.4d, 1.7c, 2.1c]

Austin, J. W. 1975. Rorty's materialism. *Auslegung* 3:20-28.

Bernstein, R. 1968. The challenge of scientific materialism. *International Philosophical Quarterly* 8:252-75.

Bush, E. 1974. Rorty revisited. *Philosophical Studies* 25:33-42.

Cam, P. 1978. "Rorty revisited", or "Rorty revised". *Philosophical Studies* 33:377-86.

Carter, W. R. 1974. On incorrigibility and eliminative materialism. *Philosophical Studies* 28:113-21.

Cornman, J. 1968. On the elimination of 'sensations' and sensations. *Review of Metaphysics* 22:15-35.

Donovan, C. 1978. Eliminative materialism reconsidered. *Canadian Journal of Philosophy* 8.

Doppelt, G. 1977. Incorrigibility, the mental, and materialism. *Philosophy Research Archives*.

Everitt, N. 1981. A problem for the eliminative materialist. *Mind* 90:428-34.

Everitt, N. 1983. How not to solve a problem for the eliminative materialist. *Mind* 92:590-92.

Feyerabend, P. 1963. Mental events and the brain. *Journal of Philosophy* 40:295-6. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Identity theory implies dualism, though its acceptance of mental properties. Instead we should eliminate talk of mental processes altogether, or redefine them in physiological terms.

Feyerabend, P. 1963. Materialism and the mind-body problem. *Review of Metaphysics* 17:49-67.

Globus, G. 1989. The strict identity theory of Schlick, Russell, Maxwell, and Feigl. In (M. Maxwell & C. Savage, eds) *Science, Mind, and Psychology: Essays in Honor of Grover Maxwell*. University Press of America.

Godow, R. 1976. Eliminative materialism and denotation. *Philosophy and Phenomenological Research* 36.

Goodman, R. B. 1974. A note on eliminative materialism. *Journal of Critical Analysis* 5:80-83.

Hiley, D. R. 1978. Is eliminative materialism materialistic? *Philosophy and Phenomenological Research* 38:325-37.

Hiley, D. R. 1980. The disappearance theory and the denotation argument. *Philosophical Studies* 37:307-20.

Lycan, W. G. & Pappas, G. 1972. What is eliminative materialism? *Australasian Journal of Philosophy* 50:149-59.

Lycan, W. G. 1976. Quine's materialism. *Philosophia* 6:101-30.

Quine, W. V. 1966. On mental entities. In *The Ways of Paradox*. Random House.

Richardson, R. C. 1981. Disappearance and the identity theory. *Canadian Journal of Philosophy* 11:473-85.

Rorty, R. 1965. Mind-body identity, privacy, and categories. *Review of Metaphysics* 19:24-54.

Rorty, R. 1970. Incorrigibility as the mark of the mental. *Journal of Philosophy*.

Rorty, R. 1970. In defense of eliminative materialism. *Review of Metaphysics* 24:112-21.

Rosenthal, D. M. 1980. Keeping matter in mind. *Midwest Studies in Philosophy* 5:295-322.

Savitt, S. 1974. Rorty's disappearance theory. *Philosophical Studies* 28:433-36.

Shirley, E. S. 1974. Rorty's "disappearance" version of the identity theory. *Philosophical Studies* 25:73-75.

Sikora, R. I. 1974. Rorty's mark of the mental and his disappearance theory. *Canadian Journal of Philosophy* 4:191-93.

Sikora, R. I. 1975. Rorty's new mark of the mental. *Analysis* 35:192-94.

Steiling, K. 1976. The elimination of sensations and the loss of philosophy. *Auslegung* 3:20-28.

3.5d Anomalous Monism (Davidson)

Davidson, D. 1970. Mental events. In (L. Foster & J. Swanson, eds) *Experience and Theory*. Humanities Press. Reprinted in *Essays on Action and Events* (Oxford University Press, 1980).

Arguing for anomalous monism: no strict psychophysical laws, no strict psychological laws, and token identity without type identity. Mental events can still cause, via subsumption under physical laws.

Davidson, D. 1973. The material mind. In (P. Suppes, ed) *Logic, Methodology and the Philosophy of Science*. North-Holland. Reprinted in *Essays on Action and Events* (Oxford University Press, 1980).

The psychological supervenes on the physical but is not reducible to it, because of the holistic nature of intentional attribution. So building a perfect physical model may not explain psychology.

Davidson, D. 1974. Psychology as philosophy. In (S. Brown, ed) *Philosophy of Psychology*. Harper & Row. Reprinted in *Essays on Action and Events* (Oxford University Press, 1980).

On the differing constitutive standards of mental and physical concepts. Attribution of mental concepts is holistic, and presupposes a background of rationality, etc. With examples from decision theory.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford University Press.

A collection of papers on action, causation and the philosophy of psychology.

Davidson, D. 1987. Problems in the explanation of action. In (P. Pettit, R. Sylvan, & J. Norman, eds) *Metaphysics and Morality*. Blackwell.

Remarks on how mental properties can explain action without strict laws. The mental is a conceptual, not an ontological category, governed by normative standards, and not reducible to the non-normative.

Davidson, D. 1992. Thinking causes. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Davidson, D. 1995. Laws and cause. *Dialectica* 49:263-79.

Antony, L. 1989. Anomalous monism and the problem of explanatory force. *Philosophical Review* 98:153-87.

Criticism of Davidson's argument for rational causation. Reasons must cause in virtue of their rational properties. Token identities can't exist, due to normativity. Quinean psychology can't yield rational explanations.

Bickle, J. 1992. Mental anomaly and the new mind-brain reductionism. *Philosophy of Science* 59:217-30.

Campbell, N. 1997. The standard objection to anomalous monism. *Australasian Journal of Philosophy* 75:373-82.

Campbell, N. 1998. Anomalous monism and the charge of epiphenomenalism. *Dialectica* 52:23-39.

Cheng, K. 1997. Davidson's action theory and epiphenomenalism. *Journal of Philosophical Research* 22:81-95.

Child, W. 1993. Anomalism, uncodifiability, and psychophysical relations. *Philosophical Review*.

Anomalism is compatible with supervenience, if it is construed as denying psychophysical laws useful for explaining behavior. It is incompatible with token identity, though. With much on the uncodifiability of rationality.

Cooper, W. E. 1980. Materialism and madness. *Philosophical Papers* 9:36-40.

Daniel, S. G. 1999. Why even Kim-style psychophysical laws are impossible. *Pacific Philosophical Quarterly* 80:225-237.

Elgin, C. 1980. Indeterminacy, underdetermination and the anomalous monism. *Synthese* 45:233-55.

Garrett, B. 1999. Davidson on causal relevance. *Ratio* 12:14-33.

Goldberg, B. 1977. A problem with anomalous monism. *Philosophical Studies* 32:175-80.

Davidson's argument equivocates on the term "physical": the physical events that mental events cause might not be subsumed under laws.

Hess, P. 1981. Actions, reasons and Humean causes. *Analysis* 41:77-81.

Anomalous monism implies that mental properties don't cause anything.

Honderich, T. 1982. The argument for anomalous monism. *Analysis* 42:59-64.

If anomalous monism is true, mental events may cause, but their mental properties aren't causally relevant.

Johnston, M. 1985. Why having a mind matters. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

Anomalous monism loses out to Australian materialism. It can't be a priori, it leads to exhaustive monism, it doesn't support a new view of free action, and it implies the causal irrelevance of the mental.

Kalderon, M. E. 1987. Epiphenomenalism and content. *Philosophical Studies* 52:71-90.

Davidson's view leads to epiphenomenalism about content, as it can't support the appropriate counterfactuals. Strong supervenience might be a way out, but that is inconsistent with anomalism.

Kernohan, A. 1985. Psychology: Autonomous or anomalous? *Dialogue* 24:427-42.

Kim, J. 1985. Psychophysical laws. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

How there can be psychophysical generalizations but no laws -- they might lack modal force. On the relation between psychophysical anomalism and psychological anomalism. Casting Davidson as a Kantian dualist.

Kim, J. 1993. Can supervenience and "non-strict laws" save anomalous monism? In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Klagge, J. C. 1990. Davidson's troubles with supervenience. *Synthese* 85:339-52.

Anomalous supervenience is consistent, at the cost of anti-realism about the mental. Supervenience is a constraint on interpretation, but needn't support counterfactuals as different interpretation schemes are possible,

Klee, R. 1992. Anomalous monism, ceteris paribus, and psychological explanation. *British Journal for the Philosophy of Science* 43:389-403.

Problems with holism and ceteris paribus laws aren't unique to psychology. One finds the same thing in the physical sciences. So rationality plays no special role, and psychological laws are as reasonable as physical laws.

Kuczynski, J. M. 1998. A proof of the partial anomalousness of the mental. *Southern Journal Of Philosophy* 36:491-504.

Latham, N. 1999. Davidson and Kim on psychophysical laws. *Synthese* 118:121-44.

LePore, E. & Loewer, B. 1987. Mind matters. *Journal of Philosophy* 630-42.

Anomalous monism is not committed to epiphenomenalism, as even non-strict laws can ground counterfactuals and so support the causal relevance of mental properties.

Lycan, W. G. 1981. Psychological laws. *Philosophical Topics* 12:9-38.

A functionalist defense against anomalous monism. Psychofunctional laws and psychological laws, though not psychophysical laws, may exist. Rebutting arguments from rationality, indeterminism, intensionality, etc.

McDowell, J. 1985. Functionalism and anomalous monism. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

Against Loar's functionalist reductionism: it doesn't begin to capture the normative role of rationality or the subjectivity of the mental.

McLaughlin, B. P. 1985. Anomalous monism and the irreducibility of the mental. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

A very thorough summary of Davidson's views. Highly recommended.

McLaughlin, B. P. & LePore, E. (eds) 1985. *Actions and Events*. Blackwell.

30 essays on Davidson.

McLaughlin, B. P. 1992. On Davidson's response to the charge of epiphenomenalism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Comments on Davidson 1992. Davidson can respond to critics accepting causal relevance of mental properties and still denying strict laws. Davidson misconstrues his critics' positions on supervenience.

Melchert, N. 1986. What's wrong with anomalous monism. *Journal of Philosophy* 80:265-74.

Davidson is concerned with intentional, not phenomenal states; and his characterization of these is just as physical states under a certain description. So he avoids epiphenomenalism (contra e.g. Honderich 1982).

Miller, A. 1993. Some anomalies in Kim's account of Davidson. *Southern Journal of Philosophy* 31:335-

44.

Kim's version of Davidson's argument against psychophysical laws cannot work. Elucidating the notion of a constitutive principle.

Noren, S. J. 1979. Anomalous monism, events, and 'the mental'. *Philosophy and Phenomenological Research* 40:64-74.

Patterson, S. A. 1996. The anomalism of psychology. *Proceedings of the Aristotelian Society* 96:37-52.

Preyer, G. 2000. Primary reasons: From radical interpretation to a pure anomalism of the mental. *Protosociology* 14:158-179.

Rosenberg, A. 1985. Davidson's unintended attack on psychology. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

Anomalous monism implies that there aren't even heteronomic psychological generalizations, as variables can't be independently measured.

Rowlands, M. 1990. Anomalism, supervenience, and Davidson on content-individuation. *Philosophia* 295-310.

Supervenience is compatible with anomalism: biconditional laws are ruled out by the disjunctive base, and the wideness of mental states rules out one-way psychophysical laws, as there's no single property in the base.

Seager, W. E. 1981. The anomalousness of the mental. *Southern Journal of Philosophy* 19:389-401.

Elucidating Davidson's argument, focusing on the argument against strict psychophysical laws. Generalizations involve disjunctive kinds and so are heteronomic and not law-like.

Seager, W. E. 1991. Disjunctive laws and supervenience. *Analysis* 51:93-98.

Argues contra Kim that supervenience is compatible with anomalous monism: the the disjunctive generalizations aren't lawlike, as they aren't confirmed by their instances.

Smart, J. J. C. 1985. Davidson's minimal materialism. In (B. Vermazen & M. Hintikka, eds) *Essays on Davidson*. Oxford University Press.

Some comments on holism, indeterminacy, anomalism, and materialism.

Smith, P. 1982. Bad news for anomalous monism? *Analysis* 42:220-4.

Response to Honderich 1982: physical events are individuated as mental states by virtue of their causal role, so the mental is causally relevant.

Sosa, E. 1993. Davidson's thinking causes. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Stanton, W. L. 1983. Supervenience and psychophysical law in anomalous monism. *Pacific Philosophical Quarterly* 64:72-9.

Supervenience entails psychophysical principles, but this is compatible with anomalous monism. On what constitutes a strict psychophysical law.

Suppes, P. 1985. Davidson's views on psychology as a science. In (B. Vermazen & M. Hintikka, eds) *Essays on Davidson*. Oxford University Press.

Various: physics is indeterministic and intensional, animals have beliefs, psychology has derived laws, and decision-theory doesn't need speech.

Tiffany, E. C. 2001. The rational character of belief and the argument for mental anomalism. *Philosophical Studies* 103:258-314.

van Gulick, R. 1980. Rationality and the anomalous nature of the mental. *Philosophy Research Archives* 7:1404.

Rationality constraints don't introduce an irreducibly normative element into intentional attributions. Rationality serves as a condition of adequacy for psychophysical theories, but it doesn't rule them out.

Vermazen, B. & Hintikka, M. (eds) 1985. *Essays on Davidson*. Oxford University Press.

12 essays on Davidson, with replies.

Walsh, D. M. 1998. Wide content individualism. *Mind* 107:625-652.

Welshon, R. 1999. Anomalous monism and epiphenomenalism. *Pacific Philosophical Quarterly* 80:103-120.

Yalowitz, S. 1997. Rationality and the argument for anomalous monism. *Philosophical Studies* 87:235-58.

Yalowitz, S. 1998. Causation in the argument for anomalous monism. *Canadian Journal of Philosophy* 28:183-226.

Zangwill, N. 1993. Supervenience and anomalous monism: Blackburn on Davidson. *Philosophical Studies* 71:59-79.

3.6 Mental Causation [see also 2.2c]

Antony, L. 1991. The causal relevance of the mental. *Mind and Language* 6:295-327.

Audi, R. 1993. Mental causation: Sustaining and dynamic. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Baker, L. R. 1993. Metaphysics and mental causation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Mental causation is incompatible with strong supervenience and causal closure of physics, as we can't distinguish high-level causes from non-causes. So reject the metaphysics and make explanation prior to causation.

Barrett, J. 1994. Rationalizing explanation and causally relevant mental properties. *Philosophical Studies* 74:77-102.

Blackburn, S. 1991. Losing your mind: Physics, identity, and folk burglar prevention. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

Arguing for the causal efficacy and scientific respectability of higher-order states, such as functional-role states. To require appeal to particular physical states is to succumb to a "Tractarian" view of physical primacy.

Block, N. 1989. Can the mind change the world? In (G. Boolos, ed) *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.

Rescuing content from epiphenomenalism via functional role argument; but then functional roles aren't really causally efficacious (cf. dormitive virtue), so epi all over again? Roles vs fillers, causation vs explanation.

Block, N. 1995. Reply: Causation and two kinds of laws. In (C. Macdonald & G. Macdonald, eds) *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford University Press.

Braun, D. 1995. Causally relevant properties. *Philosophical Perspectives* 9:447-75.

Brewer, B. 1995. Compulsion by reason (Mental Causation II). *Aristotelian Society Supplement* 69:237-53.

Burge, T. 1993. Mind-body causation and explanatory practice. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Mental causation is not a real worry, but the to-do shows that materialist metaphysics has shed little light on it. It needs to be understood at the mental level. With remarks on exclusion arguments and token identity.

Crane, T. 1990. On an alleged analogy between numbers and propositions. *Analysis* 50:224-30.

How can a relation to a proposition (an abstract object) be causally efficacious? Analogy with numbers doesn't work: weight properties are only pseudo-relational, depending on units, but propositions are absolute.

Crane, T. 1992. Mental causation and mental reality. *Proceedings of the Aristotelian Society* 66:185-202.

Argues that anomalism and causal closure don't pose problems for mental causation as they are false, and that functional properties can be efficacious. States with content may be efficacious, although content itself may not be.

Crane, T. 1995. The mental causation debate (Mental causation I). *Aristotelian Society Supplement* 69:211-36.

Argues that mental causation is a deep problem for constitutive (but not identity) forms of physicalism. The only way out is to argue that it is a different variety of causation. But then what motivates physicalism?

Dretske, F. 1993. Mental events as structuring causes of behavior. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Mental events are structuring causes of behavior; biological events are triggering causes, dependent on previous mental structuring. This allows extrinsic properties to play a causal role.

Ehring, D. 1996. Mental causation, determinables, and property instances. *Nous* 30:461-80.

Hardcastle, V. G. 1998. On the matter of minds and mental causation. *Philosophy and Phenomenological Research* 58:1-25.

Heil, J. 1992. Mentality and causality. *Topoi* 11:103-110.

On various problems with mental causation, and the relationship between psychology and philosophy.

Honderich, T. 1993. The union theory and anti-individualism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

The identity theory and psychoneural correlation can't handle mental causation; only the union theory can. Anti-individualism causes problems, but should be rejected in any case.

Horgan, T. 1989. Mental quausation. *Philosophical Perspectives* 3:47-74.

How mental events are causally relevant qua mental: via an account of "qua" causation in general, using counterfactuals on "pertinently similar worlds".

Horgan, T. 1997. Kim on mental causation and causal exclusion. *Philosophical Perspectives* 11:165-84.

Hornsby, J. 1993. Agency and causal explanation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Jackson, F. & Pettit, P. 1990. Causation and the philosophy of mind. *Philosophy and Phenomenological Research Supplement* 50:195-214.

A defense of functional role as a causally efficacious property of physical states. With application to connectionism & eliminativism.

Jackson, F. & Pettit, P. 1990. Program explanation: A general perspective. *Analysis* 50:107-17.

Jackson, F. 1995. Essentialism, mental properties, and causation. *Proceedings of the Aristotelian Society*.

How can content properties be causes, given that content is a matter of functional role and that functional properties are not causes? Defends a type-identity answer against various objections.

Jackson, F. 1996. Mental causation. *Mind* 105:377-413.

A "state of the art" review paper, concentrating on problems posed by autonomy, functionalism, and externalism, and advocating a sort of identity theory. With discussion of a "map-system" view vs. a language of thought.

Kazez, J. R. 1995. Can counterfactuals save mental causation? *Australasian Journal of Philosophy* 73:71-90.

Kim, J. 1984. Epiphenomenal and supervenient causation. *Midwest Studies in Philosophy* 9:257-70. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Psychological causation, like all macrocausation, is supervenient epiphenomenal causation.

Kim, J. 1992. The nonreductivist's trouble with mental causation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Argues that nonreductive materialism implies downward causation (as the mental has more causal powers than the physical alone), and that downward causation violates the causal closure of the physical.

Kim, J. 1992. "Downward causation" in emergentism and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Argues that nonreductive materialism is just like 1930s emergentism, with the the mental contributing new causal powers, and so implies downward causation.

Kim, J. 1993. Mental causation in a physical world. In (E. Villanueva, ed) *Science and Knowledge*. Ridgeview.

Kim, J. 1994. 'Second-order' properties and mental causation. Manuscript.

Kim, J. 1995. Mental causation: What? Me worry? In (E. Villanueva, ed) *Content*. Ridgeview.

Leiter, B. & Miller, A. 1994. Mind doesn't matter yet. *Australasian Journal of Philosophy* 72:220-28.

Argues that the arguments of Fodor and LePore & Loewer don't succeed in defeating the threat of epiphenomenalism.

Leiter, B. & Miller, A. 1998. Closet dualism and mental causation. *Canadian Journal of Philosophy* 28:161-181.

LePore, E. & Loewer, B. 1989. More on making mind matter. *Philosophical Topics* 17:175-91.

On the problems that irreducibility -- multiple realizability, normativity, and non-supervenience -- poses for mental causation. Criticizes Kim's supervenient causation and Fodor's causal powers, and looks to "quasation".

Macdonald, C. & Macdonald, G. 1986. Mental causes and explanation of action. *Philosophical Quarterly* 36:145-58.

Macdonald, C. & Macdonald, G. 1991. Mental causation and nonreductive monism. *Analysis* 51:23-32.

Macdonald, C. & Macdonald, G. 1995. How to be psychologically relevant. In (C. Macdonald & G. Macdonald, eds) *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford University Press.

Macdonald, G. 1992. The nature of naturalism. *Aristotelian Society Supplement* 66:225-44.

Marras, A. 1994. Nonreductive materialism and mental causation. *Canadian Journal of Philosophy* 24:465-93.

Marras, A. 1997. The causal relevance of mental properties. *Philosophia* 25:389-400.

Marras, A. 1998. Kim's principle of explanatory exclusion. *Australasian Journal of Philosophy* 76:439-451.

McGrath, M. 1998. Proportionality and mental causation: A fit? *Philosophical Perspectives* 12:167-176.

McLaughlin, B. P. 1989. Type epiphenomenalism, type dualism, and the causal priority of the physical. *Philosophical Perspectives* 3:109-135.

Physical comprehensiveness and mental/physical non-reductionism don't imply mental inefficacy; nor does anomalous monism. Non-physical types can still be causal, though they must be accompanied by physical causation.

Noordhof, P. 1997. Making the change: The functionalist's way. *British Journal for the Philosophy of Science* 48:233-??.

Noordhof, P. 1998. Do tropes resolve the problem of mental causation? *Philosophical Quarterly* 48:221-26.

Pettit, P. 1992. The nature of naturalism. *Aristotelian Society Supplement* 66:245-66.

On making sense of the causal efficacy of higher-level properties under naturalism. They're relevant at the program level, not quite in the way that basic properties are. With remarks on Macdonald's objections.

Robb, D. 1997. The properties of mental causation. *Philosophical Quarterly* 187:178-94.

Robb, D. 2001. Reply to Noordhof on mental causation. *Philosophical Quarterly* 51:90-94.

Robinson, W. S. 1979. Do pains make a difference to our behavior? *American Philosophical Quarterly* 16:327-34.

On Goldman's (1969) argument that dualism and causal closure are compatible with mental causation. Goldman establishes only hypothetical necessity, not causal necessity

Searle, J. R. 1984. Intentionality and its place in nature. *Synthese* 61:3-16.

Intentionality is caused by the physical, and causes. More a 1P emphasis.

Sosa, E. 1984. Mind-body interaction and supervenient causation. *Midwest Studies in Philosophy* 9:271-81.

Interactionist dualism is out, supervenient causation is in. But there are problems with mental events' causal relevance qua mental, especially for anomalous monism. Cf: a loud shot causes death, but loudness

isn't relevant.

Thomasson, A. 1998. A nonreductivist solution to mental causation. *Philosophical Studies* 89:181-95.

Tuomela, R. 1998. A defense of mental causation. *Philosophical Studies* 90:1-34.

van Gulick, R. 1993. Who's in charge here? And who's doing all the work? In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

On three arguments against mental causation, from strict laws, non-local supervenience, and especially exclusion. Mental properties are stable, recurring high-level patterns with their own causal relevance.

Worley, S. 1997. Determination and mental causation. *Erkenntnis* 46:281-304.

Yablo, S. 1992. Mental causation. *Philosophical Review* 101:245-280.

Argues that mental events/properties stand to physical events/properties as determinable to determinates, solving the exclusion problem. Some mental events are *better* candidates for the cause of action than physical events.

Zangwill, N. 1996. Good old supervenience: Mental causation on the cheap. *Synthese* 106:67-101.

Argues that anomalous monism is compatible with mental causation: supervenience is necessary and sufficient for causal efficacy.

3.7 Personal Identity

3.7a

Personal Identity, General

Baillie, J. 1993. Recent work on personal identity. *Philosophical Books* 34:193-206.

Baillie, J. 1997. Personal identity and mental content. *Philosophical Psychology* 10:323-33.

Brennan, A. 1982. Personal identity and personal survival. *Analysis* 42:44-50.

Brennan, A. 1984. Survival. *Synthese* 59:339-62.

Brennan, A. 1987. Discontinuity and identity. *Nous* 21:241-60.

- Brennan, A. 1988. *Conditions of Identity: A Study of Identity and Survival*. Oxford University Press.
- Brooks, D. H. M. 1986. Group minds. *Australasian Journal of Philosophy* 64:456-70.
- Carter, W. 1999. Will I be a dead person? *Philosophy and Phenomenological Research* 59.
- Cartwright, H. M. 1987. Ruminations on an account of personal identity. In (J. J. Thomson, ed) *On Being and Saying: Essays on Honor of Richard Cartwright*. MIT Press.
- Cartwright, H. M. 1993. On two arguments for the indeterminacy of personal identity. *Synthese* 95:241-273.
- Cockburn, D. (ed) 1991. *Human Beings*. Cambridge University Press.
- Coleman, S. 2000. Thought experiments and personal identity. *Philosophical Studies* 98:51-66.
- Cowley, F. 1971. The identity of a person and his body. *Journal of Philosophy* 68:678-683.
- Dainton, B. 1996. Survival and experience. *Proceedings of the Aristotelian Society* 96:17-36.
- Davis, L. H. 1998. Functionalism and personal identity. *Philosophy and Phenomenological Research* 58:781-804.
- Davis, L. H. 2001. Functionalism, the brain, and personal identity. *Philosophical Studies* 102:259-79.
- Dennett, D. C. 1978. Where am I? In *Brainstorms*. MIT Press.
- Elliot, R. 1991. Personal identity and the causal continuity requirement. *Philosophical Quarterly* 41:55-75.
- Ganeri, J. 2000. Cross-modality and the self. *Philosophy and Phenomenological Research* 61:639-658.
- Garrett B. 1990. Personal identity and extrinsicness. *Philosophical Studies* 59:177-194.
- Garrett, B. 1991. Personal identity and reductionism. *Philosophy and Phenomenological Research* 51:361-373.
- Garrett, B. 1992. Persons and values. *Philosophical Quarterly* 42:337-44.
- Glover, J. 1988. *I: The Philosophy and Psychology of Personal Identity*. Penguin.
- Hamilton, A. 1995. A new look at personal identity. *Philosophical Quarterly* 45:332-349.

- Harris, H. (ed) 1995. *Identity*. Oxford University Press.
- Harris, H. 1995. An experimentalist looks at identity. In (H. Harris, ed) *Identity*. Oxford University Press.
- Hasker, W. 1999. *The Emergent Self*. Cornell University Press.
- Hope, T. 1994. Personal Identity and Psychiatric Illness. *Philosophy* 37:131-143.
- Johnston, M. 1992. Reasons and reductionism. *Philosophical Review* 3:589-618.
- Kolak, D. & Martin, R. 1987. Personal identity and causality: Becoming unglued. *American Philosophical Quarterly*.
- Kolak, D. 1993. The metaphysics and metapsychology of personal identity: Why thought experiments matter in deciding who we are. *American Philosophical Quarterly* 30:39-50.
- Kolak, D. & Martin, R. (eds) 1991. *Self and Identity: Contemporary Philosophical Issues*. Macmillan.
- Madell, G. 1981. *The Identity of the Self*. Edinburgh University Press.
- Madell, G. 1991. Personal identity and the idea of a human being. *Philosophy* 29:127-142.
- Martin, R. 1992. Self-interest and survival. *American Philosophical Quarterly* 29:319-30.
- Matthews, S. 2000. Survival and separation. *Philosophical Studies* 98:279-303.
- McCall, C. 1990. *Concepts of Person: An Analysis of Concepts of Person, Self, and Human Being*. Avebury.
- Merricks, T. 2000. Perdurantism and psychological continuity. *Philosophy and Phenomenological Research* 61:195-199.
- Miri, M. 1973. Memory and personal identity. *Mind* 82:1-21.
- Nerlich, G. C. 1958. Sameness, difference, and continuity. *Analysis*.
- Noonan, H. 1989. *Personal Identity*. Routledge.
- Noonan, H. 1993. Chisholm, persons, and identity. *Philosophical Studies* 69:35-58.

- Nozick, R. 1981. The identity of the self. In *Philosophical Explanations*. Harvard University Press.
- Olson E. 1994. Is Psychology relevant to personal identity? *Australasian Journal of Philosophy* 72:173-186.
- Olson, E. T. 1997. *The Human Animal: Personal Identity without Psychology*. Oxford University Press.
- Olson, E. 2001. Personal identity and the radiation argument. *Analysis* 61:38-44.
- Peacocke, A. & Gillett, G. (eds) 1987. *Persons and Personality: A Contemporary Inquiry*. Blackwell.
- Penelhum, T. 1959. Personal identity, memory, and survival. *Journal of Philosophy*.
- Penelhum, T. 1971. The importance of self-identity. *Journal of Philosophy* 68:667-78.
- Perry, J. 1972. Can the self divide? *Journal of Philosophy* 69:463-88.
- Perry, J. (ed) 1975. *Personal Identity*. University of California Press.
- Perry, J. 1975. Personal identity, memory, and the problem of circularity. In (J. Perry, ed) *Personal Identity*. University of California Press.
- Perry, J. 1976. The importance of being identical. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Perry, J. 1978. *A Dialogue on Personal Identity and Immortality*. Hackett.
- Pogue, J. E. 1993. Identity, survival, and the reasonableness of replication. *Southern Journal of Philosophy* 31:45-70.
- Rea, M. & Silver, D. 2000. Personal identity and psychological continuity. *Philosophy and Phenomenological Research* 61:185-194.
- Rey, G. 1976. Survival. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Rieber, S. 1998. The concept of personal identity. *Philosophy and Phenomenological Research* 58:581-594.
- Robert, M. 1983. Lewis's theory of personal identity. *Australasian Journal of Philosophy* 61:58-67.
- Rorty, A. (ed) 1976. *The Identities of Persons*. University of California Press.

- Shalom, A. 1985. *The Body-Mind Conceptual Framework and the Problem of Personal Identity*. Humanities Press.
- Schechtman, M. 1990. Personhood and personal identity. *Journal of Philosophy* 87:71-92.
- Shoemaker, S. 1959. Personal identity and memory. *Journal of Philosophy* 56:868-902.
- Shoemaker, S. 1970. Persons and their pasts. *American Philosophical Quarterly* 7:269-85.
- Shoemaker, S. & Swinburne, S. 1984. *Personal Identity: Great Debates in Philosophy*. Blackwell.
- Shorter, J. M. 1962. More about bodily continuity and personal identity. *Analysis* 22:79-85.
- Unger, P. 1990. *Identity, Consciousness, and Value*. Oxford University Press.
- Vesey, P. 1974. *Personal Identity: A Philosophical Analysis*. Cornell University Press.
- White, S. 1989. Metapsychological relativism and the self. *Journal of Philosophy* 86:298-323.
- Whiting, J. 1986. Friends and future selves. *Philosophical Review* 95:547-80.
- Wilkes, K. V. 1988. *Real People: Personal Identity Without Thought Experiments*. Oxford University Press.
- Williams, B. 1957. Personal identity and individuation. *Proceedings of the Aristotelian Society* 67:229-52.
- Williams, B. 1973. *Problems of the Self*. Cambridge University Press.
- Zemach, E. 1987. Looking out for number one. *Philosophy and Phenomenological Research*.
- Zuboff, A. 1978. Moment universals and personal identity. *Proceedings of the Aristotelian Society* 52:141-55.
- Zuboff, A. 1990. One self: The logic of experience. *Inquiry* 33:39-68.

3.7b

Parfit on Personal Identity

- Baillie, J. 1993. What matters in survival. *Southern Journal of Philosophy* 31:255-61.
- Baillie, J. 1996. Identity, relation R, and what matters: A challenge to Derek Parfit. *Pacific Philosophical Quarterly* 77:263-267.
- Beck, S. 1989. Parfit and the Russians (personal identity and moral concepts). *Analysis* 49:205-209.
- Bodansky, E. 1987. Parfit on selves and their interests. *Analysis* 47:47-50.
- Brennan, A. A. 1987. Survival and importance. *Analysis* 47:225-30.
- Brueckner, A. 1993. Parfit on what matters in survival. *Philosophical Studies* 70:1-22.
- Bushnell, D. E. 1993. Identity, psychological continuity, and rationality. *Journal of Philosophical Research* 18:15-24.
- Campbell, S. 2000. Strawson, Parfit and impersonality. *Canadian Journal of Philosophy* 30:207-225.
- Cassam, Q. 1993. Parfit on persons. *Proceedings of the Aristotelian Society* 93:17-37.
- Chappell, T. 1995. Personal identity, R-relatedness, and the empty question argument. *Philosophical Quarterly* 45:88-92.
- Chappell, T. 1998. Reductionism about persons; and what matters. *Proceedings of the Aristotelian Society* 98:41-58.
- Collins, A. W. 1997. Personal identity and the coherence of q-memory. *Philosophical Quarterly* 47:73-80.
- Curzer, H. 1991. An ambiguity in Parfit's theory of personal identity. *Ratio* 4:16-24.
- Dancy, J. (ed). 1997. *Reading Parfit*. Blackwell.
- Doepke, F. 1990. The practical importance of personal identity. *Logos* 83-91.
- Ehring, D. 1987. Survival and trivial facts. *Analysis* 47:50-54.
- Ehring, D. 1995. Personal identity and the R-relation: Reconciliation through cohabitation. *Australasian Journal of Philosophy* 73:337-346.
- Fields, L. 1987. Parfit on personal identity and desert. *Philosophical Quarterly* 37:432-41.

- Gillett, G. 1987. Reasoning about persons. In (A. Peacocke & G. Gillett, eds) *Persons and Personality: A Contemporary Inquiry*. Blackwell.
- Goodenough, J. M. 1996. Parfit and the Sorites paradox. *Philosophical Studies* 2:113-20.
- Haugen, D. 1995. Personal identity and concern for the future. *Philosophia* 24:481-492.
- Hirsch, E. 1991. Divided minds. *Philosophical Review* 1:3-30.
- Johnston, M. 1989. Fission and the facts. *Philosophical Perspectives* 3:369-97.
- Korsgaard, C. 1989. Personal identity and the unity of agency: A Kantian response to Parfit. *Philosophy and Public Affairs* 18:103-31.
- Lee, W. 1990. Personal identity, the temporality of agency, and moral responsibility. *Auslegung* 16:17-29.
- Lewis, D. 1976. Survival and identity. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Madell, G. 1985. Derek Parfit and Greta Garbo. *Analysis* 45:105-9.
- Maddy, P. 1979. Is the importance of identity derivative? *Philosophical Studies* 35:151-70.
- Matthews, G. B. 1977. Surviving as. *Analysis* 37:53-58.
- Martin, R. 1987. Memory, connecting, and what matters in survival. *Australasian Journal of Philosophy* 65:82-97.
- Measor, N. 1980. On what matters in survival. *Mind* 89:406-11.
- Merricks, T. 1997. Fission and personal identity over time. *Philosophical Studies* 88:163-186.
- Northoff, G. 2000. Are "q-memories" empirically realistic?: A neurophilosophical approach. *Philosophical Psychology* 13:191-211.
- Oaklander, L. N. 1987. Parfit, circularity, and the unity of consciousness. *Mind* 96:525-29.
- Parfit, D. 1971. Personal identity. *Philosophical Review* 80:3-27.
- Parfit, D. 1971. On the importance of self-identity. *Journal of Philosophy* 68:683-90.

- Parfit, D. 1973. Later selves and moral principles. In (A. Montefiore, ed) *Philosophy and Personal Relations*. Routledge and Kegan Paul.
- Parfit, D. 1976. Lewis, Perry, and what matters. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Parfit, D. 1982. Personal identity and rationality. *Synthese* 53.
- Parfit, D. 1984. *Reasons and Persons*. Oxford University Press.
- Parfit, D. 1995. The unimportance of identity. In (H. Harris, ed) *Identity*. Oxford University Press.
- Robinson, J. 1988. Personal identity and survival. *Journal of Philosophy* 85:319-28.
- Rovane, C. 1990. Branching self-consciousness. *Philosophical Review* 99:355-95.
- Siderits, M. 1988. Ehrling on Parfit's relation R. *Analysis* 48:29-32.
- Slors, M. 2001. Personal identity, memory, and circularity: An alternative for q-memory. *Journal of Philosophy* 98:186-214.
- Sprigge, T. L. S. 1988. Personal and impersonal identity. *Mind* 97:29-49.
- Storl, H. 1992. The problematic nature of parfitian persons. *Personalist Forum* 8:123-31.
- Stone, J. 1988. Parfit and the Buddha: Why there are no people. *Philosophy and Phenomenological Research* 48:519-32.
- Wolf, S. 1986. Self-interest and interest in selves. *Ethics* 96:704-20.

3.7c

Persons

- Aune, B. 1994. Speaking of selves. *Philosophical Quarterly* 44:279-93.
- Barresi, J. 1999. On becoming a person. *Philosophical Psychology* 12:79-98.
- Bertocci, P. A. 1978. The essence of a person. *Monist* 61:28-41.
- Biro, J. I. 1981. Persons as corporate entities and corporations as persons. *Nature and System* 3:173-80.

- Chisholm, R. M. 1976. *Person and Object: A Metaphysical Study*. Open Court.
- Dennett, D. C. 1976. Conditions of personhood. Lewis, D. 1976. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Dennett, D. C. 1989. The origins of selves. *Cogito* 3:163-73.
- Heinimaa, M. 2000. Ambiguities in the psychiatric use of the concepts of the person: An analysis. *Philosophy, Psychiatry, and Psychology* 7:125-136.
- Lowe, E. J. 1991. Real selves: Persons as a substantial kind. *Philosophy* 29:87-107.
- Johnston, M. 1987. Human beings. *Journal of Philosophy* 84:59-83.
- Margolis, J. 1988. Minds, selves, and persons. *Topoi* 7:31-45.
- McInerney, P. K. 1998. Persons and psychological systems. *American Philosophical Quarterly* 35:179-193.
- McInerney, P. K. 2000. Conceptions of persons and persons through time. *American Philosophical Quarterly* 37:121-134.
- Oderberg, D. 1989. Johnston on human beings. *Journal of Philosophy* 86:137-41.
- Olson, E. 1998. Human atoms. *Australasian Journal of Philosophy* 76:396-406.
- Peterson, J. 1985. Persons and the problem of interaction. *Modern Schoolman* 62:131-38.
- Rorty, A. O. 1976. A literary postscript: Characters, persons, selves, individuals. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Smart, B. 1976. Synchronous and diachronous selves. *Canadian Journal of Philosophy* 6:13-33.
- Strawson, P. 1958. Persons. *Minnesota Studies in the Philosophy of Science* 2:330-53.
- Unger, P. 1979. I do not exist. In (G. Macdonald, ed) *Perception and Identity*. Cornell University Press.
- Unger, P. 1979. Why there are no people. *Midwest Studies in Philosophy* 4:177-222.
- Vincent, A. 1989. Can groups be persons? *Review of Metaphysics* 42:687-715.

Wiggins, D. 1987. The person as object of science, as subject of experience, and as locus of value. In (A. Peacocke & G. Gillett, eds) *Persons and Personality*. Blackwell.

3.7d

Split Brains [see also 6.1e]

Baillie, J. 1991. Split brains and single minds. *Journal of Philosophical Research* 16:11-18.

Davis, L. 1997. Cerebral hemispheres. *Philosophical Studies* 87:207-22.

Gill, J. H. 1980. Of split brains and tacit knowing. *International Philosophical Quarterly* 20:49-58.

Gillett, G. 1986. Brain bisection and personal identity. *Mind* 95:224-9.

Greenwood, J. D. 1993. Split brains and singular personhood. *Southern Journal of Philosophy* 31:285-306.

Marks, C. 1980. *Commissurotomy, Consciousness, and Unity of Mind*. MIT Press.

Martin, R. 1995. Fission rejuvenation. *Philosophical Studies* 80:17-40.

Merricks, T. 1997. Fission and personal identity over time. *Philosophical Studies* 88:163-186.

Mills E. 1993. Dividing without reducing: Bodily fission and personal identity. *Mind* 102:37-51.

Moor, J. 1982. Split brains and atomic persons. *Philosophy of Science* 49:91-106.

Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22:396-413. Reprinted in *Mortal Questions* (Cambridge University Press, 1979).

Parfit, D. 1987. Divided minds and the nature of persons. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Puccetti, R. 1973. Brain bisection and personal identity. *British Journal for the Philosophy of Science* 24:339-55.

Puccetti, R. 1973. Multiple identity. *Personalist* 54:203-13.

Puccetti, R. 1975. The mute self: A reaction to DeWitt's alternative account of the split-brain data. *British*

Journal for the Philosophy of Science 27:65-73.

Puccetti, R. 1981. The case for mental duality: Evidence from split-brain data and other considerations. Behavioral and Brain Sciences 4:93-123.

Puccetti, R. 1989. Two brains, two minds. British Journal for the Philosophy of Science 40:137-44.

Puccetti, R. 1993. Mind with a double brain. British Journal for the Philosophy of Science 44:675-92.

Puccetti, R. 1993. Dennett on the split-brain. Psycholoquy 4(52).

Robinson, D. N. 1976. What sort of persons are hemispheres? Another look at "split-brain" man. British Journal for the Philosophy of Science 27:73-8.

Shaffer, J. 1977. Personal identity: The implications of brain bisection and brain transplants. Journal of Medicine and Philosophy 2:147-61.

Sperry, R. W. 1984. Consciousness, personal identity and the divided brain. Neuropsychologia 22:611-73.

3.7e Multiple Personality

Apter, A. 1991. The problem of who: Multiple personality, personal identity, and the double brain. Philosophical Psychology 4:219-48.

Benner, D. G. , Evans, C. S. 1984. Unity and multiplicity in hypnosis, commissurotomy, and multiple personality disorder. Journal of Mind and Behavior 5:423-431.

Boden, M. A. 1994. Multiple personality and computational models. Philosophy 37:103-114.

Braude, S. E. 1991. *First-person Plural: Multiple Personality and the Philosophy of Mind*. Routledge.

Braude, S. E. 1996. Multiple personality disorder and moral responsibility. Philosophy, Psychiatry, and Psychology 3:37-54.

Clark, S. R. L. 1991. How many selves make me? Philosophy 29:213-33.

Clark, S. R. L. 1996. Minds, memes, and multiples. Philosophy, Psychiatry, and Psychology 3:21-28.

Flanagan, O. 1994. Multiple identity, character transformation, and self-reclamation. In (G. Graham & G. Stephens, eds) *Philosophical Psychopathology*. MIT Press.

- Gillett, G. 1986. Multiple personality and the concept of a person. *New Ideas in Psychology* 4:173-84.
- Gillett, G. 1997. A discursive account of multiple personality disorder. *Philosophy, Psychiatry, and Psychology* 4:213-22.
- Hacking, I. 1991. Two souls in one body. *Critical Inquiry* 17:838-67.
- Hacking, I. 1995. *Rewriting the Soul: Multiple Personality and the Sciences of Memory*. Princeton University Press.
- Humphrey, N. & Dennett, D. C. 1989. Speaking for ourselves. *Raritan* 9:68-98.
- Kolak, D. 1993. Finding our selves: Identification, identity, and multiple personality. *Philosophical Psychology* 6:363-86.
- Lizza, J. P. 1993. Multiple personality and personal identity revisited. *British Journal for the Philosophy of Science* 44:263-274.
- Matthews, S. 1998. Personal identity, multiple personality disorder, and moral personhood. *Philosophical Psychology* 11:67-88.
- Radden, J. 1996. *Divided Minds and Successive Selves: Ethical Issues in Disorders of Identity and Personality*. MIT Press.
- Wilkes, K. V. 1981. Multiple personalty and personal identity. *British Journal for the Philosophy of Science* 32:331-48.
- Wilkes, K. V. 1991. How many selves make me? *Philosophy* 66:235-43.

3.8

Free Will

- Albritton, R. 1985. Freedom of the will and freedom of action. *Proceedings and Addresses of the American Philosophical Association* 59:239-51.
- Anscombe, G. E. M. 1976. 'Soft' determinism. In (G. Ryle, ed) *Contemporary Aspects of Philosophy*. Oriel Press.
- Audi, R. 1974. Moral responsibility, freedom, and compulsion. *American Philosophical Quarterly* 11:1-14.

- Ayer, A. J. 1980. Free will and rationality. In (Z. van Straaten, ed) *Philosophical Subjects*. Oxford University Press.
- Ayers, M. 1968. *The Refutation of Determinism*. Methuen.
- Berofsky, B. (ed) 1966. *Free Will and Determinism*. Harper and Row.
- Berofsky, B. 1971. *Determinism*. Princeton University Press.
- Berofsky, B. 1987. *Freedom from Necessity: The Metaphysical Basis of Responsibility*. Routledge.
- Bishop, J. 1993. Compatibilism and the free will defense. *Australasian Journal of Philosophy* 71:104-20.
- Blumenfeld, D. 1971. The principle of alternate possibilities. *Journal of Philosophy* 67:339-44.
- Blumenfeld, D. 1988. Freedom and mind control. *American Philosophical Quarterly* 25:215-27.
- Campbell, C. A. 1951. Is "free will" a pseudoproblem? *Mind* 60:441-65.
- Churchland, P. S. 1981. Is determinism self-refuting? *Mind* 90:99-101.
- Clarke, R. 1992. Free will and the conditions of moral responsibility. *Philosophical Studies* 66:53-72.
- Clarke, R. 1993. Toward a credible agent-causal account of free will. *Nous* 27:191-203.
- Clarke, R. 2000. Modest libertarianism. *Philosophical Perspectives* 14:21-46.
- Crisp, T. & Warfield, T. 2000. The irrelevance of indeterministic counterexamples to principle beta. *Philosophy & Phenomenological Research* 61:173-185.
- Dennett, D. C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.
- Double, R. 1989. Puppeteers, hypnotists, and neurosurgeons. *Philosophical Studies* 56:163-73.
- Double, R. 1991. *The Non-Reality of Free Will*. Oxford University Press.
- Double, R. 1991. Determinism and the experience of freedom. *Pacific Philosophical Quarterly* 72:1-8.
- Double, R. 1992. How rational must free will be? *Metaphilosophy* 23:268-78.

- Double, R. 1994. How to frame the free will problem. *Philosophical Studies* 75:149-72.
- Double, R. 1996. *Metaphilosophy and Free Will*. Oxford University Press.
- Duggan, T. & Gert, B. 1979. Free will as the ability to will. *Nous* 13:197-217.
- Dworkin, G. (ed) 1970. *Determinism, Free Will, and Moral Responsibility*. Prentice-Hall.
- Eccles, J. 1976. Brain and free will. In (G. Globus, ed) *Consciousness and the Brain*. Plenum Press.
- Fischer, J. M. 1982. Responsibility and control. *Journal of Philosophy* 79:24-40.
- Fischer, J. M. & Ravizza, M. 1992. When the will is free. *Philosophical Perspectives* 6:423-51.
- Fischer, J. M. 1994. *The Metaphysics of Free Will*. Blackwell.
- Fischer, J. M. & Ravizza, M. 1996. Free will and the modal principle. *Philosophical Studies* 3:213-30.
- Fowler, C. 1996. A pragmatic defense of free will. *Journal of Value Inquiry* 30:247-60.
- Frankfurt, H. 1969. Alternate possibilities and moral responsibility. *Journal of Philosophy* 65:829-39.
- Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68:5-20.
- Furlong, F. W. 1981. Determinism and free will: Review of the literature. *American Journal of Psychiatry* 138:435-39.
- Garson, J. W. 1995. Chaos and free will. *Philosophical Psychology* 8:365-74.
- Ginet, C. 1980. The conditional analysis of freedom. In (P. van Inwagen, ed) *Time and Cause: Essays Presented to Richard Taylor*. Reidel.
- Ginet, C. 1983. In defense of incompatibilism. *Philosophical Studies* 44:391-400.
- Ginet, C. 1989. Reasons explanation of action: An incompatibilist account. *Philosophical Perspectives* 3.
- Ginet, C. 1990. *On Action*. Cambridge University Press.
- Goldman, A. 1968. Actions, predictions, and books of life. *American Philosophical Quarterly*.
- Goldman, A. 1969. The compatibility of mechanism and purpose. *Philosophical Review* 78:468-82.

- Griffiths, A. P. 1989. Is free will incompatible with something or other? *Philosophy* 24:101-19.
- Greenspan, P. S. 1978. Behavior control and freedom of action. *Philosophical Review* 87:225-40.
- Greenspan, P. S. 1993. Free will and the genome project. *Philosophy and Public Affairs* 22:31-43.
- Heller, M. 1996. The mad scientist meets the robot cats: Compatibilism, kinds, and counterexamples. *Philosophy and Phenomenological Research* 56:333-37.
- Honderich, T. (ed) 1973. *Essays on Freedom of Action*. Routledge and Kegan Paul.
- Honderich, T. 1988. *A Theory of Determinism*. Oxford University Press.
- Honderich, T. 1993. *How Free Are You?* Oxford University Press.
- Horgan, T. 1985. Compatibilism and the consequence argument. *Philosophical Studies* 47:339-56.
- Hospers, J. 1950. Meaning and free will. *Philosophy and Phenomenological Research* 10:307-30.
- Howard, G. S. 1993. Steps toward a science of free will. *Counseling and Values* 37:116-28.
- Kane, R. 1985. *Free Will and Values*. SUNY Press.
- Kane, R. 1989. Two kinds of incompatibilism. *Philosophy and Phenomenological Research* 69:219-54.
- Kane, R. 1994. Free will: The illusive ideal. *Philosophical Studies* 75:25-60.
- Kane, R. 1996. *The Significance of Free Will*. Oxford University Press.
- Kane, R. 2000. The dual regress of free will and the role of alternative possibilities. *Philosophical Perspectives* 14:57-80.
- Kapitan, T. 1986. Deliberation and the presumption of open alternatives. *Southern Journal of Philosophy* 40:230-51.
- Kapitan, T. 1991. Ability and cognition: A defense of compatibilism. *Philosophical Studies* 63:231-43.
- Kapitan, T. 1996. Modal principles in the metaphysics of free will. *Philosophical Perspectives* 10:419-45.
- Kapitan, T. 2000. Autonomy and manipulated freedom. *Philosophical Perspectives* 14:81-104.

- Kenny, A. 1976. *Will, Freedom, and Power*. Blackwell.
- Kenny, A. 1978. *Free Will and Responsibility*. Routledge.
- Klein, M. 1990. Determinism, blameworthiness, and deprivation. Oxford University Press.
- Ladd, J. 1952. Free will and voluntary action. *Philosophy and Phenomenological Research* 12:392-405.
- Lahav, R. 1991. Between pre-determinism and arbitrariness: A Bergsonian approach to free will. *Southern Journal of Philosophy* 29:487-99.
- Lamb, J. W. 1977. On a proof of incompatibilism. *Philosophical Review* 86:20-35.
- Lamb, J. W. 1993. Evaluative compatibilism and the principle of alternate possibilities. *Journal of Philosophy* 90:517-27.
- Lehrer, K. (ed) 1966. *Freedom and Determinism*. Random House.
- Lehrer, K. 1966. An empirical disproof of determinism. In (K. Lehrer, ed) *Freedom and Determinism*. Random House.
- Lehrer, K. 1976. 'Can' in theory and practice: A possible worlds analysis. In (M. Brand & D. Walton, eds) *Action Theory*. Reidel.
- Locke, D. 1975. Three concepts of free action. *Aristotelian Society Supplement* 75:95-112.
- Lucas, J. R. 1970. *The Freedom of the Will*. Oxford University Press.
- Machina, K. 1994. Challenges for compatibilism. *American Philosophical Quarterly* 31:213-22.
- Margenau, H. 1931. The uncertainty principle and free will. *Science*.
- McCall, S. 1984. Freedom defined as the power to decide. *American Philosophical Quarterly* 21:329-38.
- Morden, M. 1990. Free will, self-causation, and strange loops. *Australasian Journal of Philosophy* 68:59-73.
- Morgenbesser, S. & Walsh, J. J. (eds) 1962. *Freedom and Responsibility*. Prentice-Hall.
- Narveson, J. 1977. Compatibilism defended. *Philosophical Studies* 32:83-7.

Neely, W. 1974. Freedom and desire. *Philosophical Review* 83:32-54.

O'Connor, D. J. 1971. *Free Will*. Anchor Books.

O'Connor, T. 1993. Indeterminism and free agency: Three recent views. *Philosophy and Phenomenological Research* 53:499-26.

O'Connor, T. (ed) 1995. *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. Oxford University Press.

O'Leary-Hawthorne, J. & Pettit, P. 1996. Strategies for free will compatibilists. *Analysis* 56:191-201.

O'Shaughnessy, B. 1980. *The Will: A Dual Aspect Theory*. Cambridge University Press.

Pereboom, D. 2000. Alternative possibilities and causal histories. *Philosophical Perspectives* 14:119-138.

Perszyk, K. J. 1999. Compatibilism and the free will defence: A reply to Bishop. *Australasian Journal of Philosophy* 77:92-105.

Popper, K. 1983. Is determinism self-refuting? *Mind* 92:103-4.

Ravizza, M. 1994. Semi-compatibilism and the transfer of non-responsibility. *Philosophical Studies* 75:61-93.

Rowe, W. 1987. Two concepts of freedom. *Proceedings and Addresses of the American Philosophical Association* 61:43-64.

Rychlak, J. F. 1976. Can psychology be objective about free will? *Philosophical Psychologist* 10:2-9. Revised version in *New Ideas in Psychology* 1:213-29, 1983.

Rychlak, J. F. 1994. Four kinds of determinism and "free will": A response to Viney and Crosby. *New Ideas in Psychology* 12:143-46.

Rychlak, J. F. 1994. Is free will a process or a content: Both? neither? Are we free to take a position on this question? *Journal of Theoretical and Philosophical Psychology* 14:62-72.

Sappington, A. A. 1990. Recent psychological approaches to the free will versus determinism controversy. *Psychological Bulletin* 108:19-29.

Searle, J. 2000. Consciousness, free action and the brain. *Journal of Consciousness Studies* 7:3-22.

- Settle, T. 1993. How determinism refutes compatibilism. *Religious Studies* 29:353-62.
- Slife, B. D. 1994. Free will and time: That "stuck" feeling. *Journal of Theoretical and Philosophical Psychology* 14:1-12.
- Slote, M. A. 1969. Free will, determinism, and the theory of important criteria. *Inquiry* 12:317-38.
- Slote, M. 1980. Understanding free will. *Journal of Philosophy* 77:136-51.
- Slote, M. 1982. Selective necessity and the free will problem. *Journal of Philosophy* 74:5-24.
- Spence, S. A. 1996. Free will in the light of neuropsychiatry. *Philosophy, Psychiatry, and Psychology* 3:75-90.
- Stampe, D. W. & Gibson, M. I. 1992. Of one's own free will. *Philosophy and Phenomenological Research* 52:529-56.
- Strawson, G. 1986. *Freedom and Belief*. Oxford University Press.
- Stump, E. & Fischer, J. 2000. Transfer principles and moral responsibility. *Philosophical Perspectives* 14:47-56.
- Thorp, J. 1980. *Free Will: A Defense Against Neurophysiological Determinism*. Routledge.
- van Inwagen, P. 1975. The incompatibility of free will and determinism. *Philosophical Studies* 27:185-99.
- van Inwagen, P. 1978. Ability and responsibility. *Philosophical Review* 87:201-24.
- van Inwagen, P. 1983. *An Essay on Free Will*. Oxford University Press.
- van Inwagen, P. 1989. When is the will free? *Philosophical Perspectives* 3.
- van Inwagen, P. 1994. When the will is not free. *Philosophical Studies* 75:95-113.
- Van Inwagen, P. 2000. Free will remains a mystery. *Philosophical Perspectives* 14:1-20.
- Vesey, G. 1989. Responsibility and free will. *Philosophy* 24:85-100.
- Vihvelin, K. 2000. Freedom, foreknowledge, and the principle of alternate possibilities. *Canadian Journal of Philosophy* 30:1-23.

- Vihvelin, K. 2000. Libertarian compatibilism. *Philosophical Perspectives* 14:139-166.
- Viney, D. W. & Crosby, D. A. 1994. Free will in process perspective. *New Ideas in Psychology* 12:129-41.
- Waller, B. N. 1989. Uneven starts and just deserts (fatalism and free will). *Analysis* 49:209-13.
- Waller, B. 1990. *Freedom without Responsibility*. Temple University Press.
- Warfield, T. 2000. Causal determinism and human freedom are incompatible: A new argument for incompatibilism. *Philosophical Perspectives* 14:167-180.
- Watson, G. (ed) 1982. *Free Will*. Oxford University Press.
- Westcott, M. R. 1977. Free will: An exercise in metaphysical truth or psychological consequences. *Canadian Psychological Review* 18:249-63.
- Widerker, D. 2000. Frankfurt's attack on the principle of alternative possibilities: A further look. *Philosophical Perspectives* 14:181-202.
- Williams, C. 1980. *Free Will and Determinism: A Dialogue*. Hackett.
- Wilton, R. 2000. *Consciousness, Free Will, and the Explanation of Human Behavior*. E. Mellen Press.
- Wolf, S. 1980. Asymmetrical freedom. *Journal of Philosophy* 77:151-66.
- Wolf, S. 1981. The importance of free will. *Mind* 90:366-78.
- Wolf, S. 1990. *Freedom within Reason*. Oxford University Press.
- Yaffe, G. 2000. Free will and agency at its best. *Philosophical Perspectives* 14:203-230.
- Zagzebski, L. 2000. Does libertarian freedom require alternate possibilities? *Philosophical Perspectives* 14:231-248.
- Zimmerman, D. 1994. Acts, omissions, and semi-compatibilism. *Philosophical Studies* 73:209-23.

3.9

The Problem of Other Minds

- Alexander, P. 1959. Other people's experiences. *Proceedings of the Aristotelian Society*.
- Aune, B. 1961. The problem of other minds. *Philosophical Review*.
- Austin, J. 1946. Other minds. *Aristotelian Society Supplement* 20:148-87.
- Ayer, A. J. 1953. One's knowledge of other minds. *Theoria*.
- Ayer, A. J. 1956. *The Problem of Knowledge*. Harmondsworth.
- Baron-Cohen, S. , Tager-Flusberg, H. , Cohen, D. J. 1994. *Understanding Other Minds: Perspectives from Autism*. Oxford University Press.
- Buck, R. 1962. Non-other minds. In (R. Butler, ed) *Analytic Philosophy*. Barnes and Noble.
- Buford, T. O. 1970. *Essays on Other Minds*. University of Illinois Press.
- Castaneda, H. 1962. Criteria, analogy, and knowledge of other minds. *Journal of Philosophy*.
- Duhrssen, A. 1963. Philosophic alienation and the problem of other minds. *Philosophical Review*.
- Everett, T. 2000. Other voices, other minds. *Australasian Journal of Philosophy* 78:213-222.
- Feigl, H. 1959. Other minds and the egocentric predicament. *Journal of Philosophy* 56:980-87.
- Gallagher, K. 1964. Intersubjective knowledge. In (Sheed & Ward, eds) *The Philosophy of Knowledge*.
- Glennan, S. S. 1995. Computationalism and the problem of other minds. *Philosophical Psychology* 8:375-88.
- Hampshire, S. 1952. The analogy of feeling. *Mind* 61:1-12.
- Harnad, S. 1991. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* 1:43-54.
- On the Total Turing Test (full behavioral equivalence) as a test for mind.
- Hyslop, A. 1976. Other minds as theoretical entities. *Australasian Journal of Philosophy* 54:158-61.
- Hyslop, A. 1995. *Other Minds*. Kluwer.

- Jones, J. R. 1950. Our knowledge of other persons. *Philosophy* 25.
- Jorgensen, J. 1949. Remarks concerning the concept of mind and the problem of other people's minds. *Theoria*.
- Kurthen, M. Moskopp, D. , Linke, D. B. & Reuter, B. M. 1991. The locked-in syndrome and the behaviorist epistemology of other minds. *Theoretical Medicine* 12:69-79.
- Lenman, J. 1994. Beliefs about other minds: A pragmatic justification. *American Philosophical Quarterly* 31:223-34.
- Locke, D. 1968. *Myself and Others: A Study in our Knowledge of Minds*. Oxford University Press.
- Malcolm, N. 1958. Knowledge of other minds. *Journal of Philosophy*.
- Melnyk, A. 1994. Inference to the best explanation and other minds. *Australasian Journal of Philosophy* 4:482-91.
- Mellor, W. W. 1956. Three problems about other minds. *Mind* 65:200-217.
- Morick, H. (ed) 1967. *Wittgenstein and the Problem of Other Minds*. Humanities Press.
- Narveson, A. H. 1966. Evidential necessity and other minds. *Mind* 75.
- Pap, A. 1951. Other minds and the principle of verifiability. *Revue Internationale de Philosophie* 5:280-306.
- Peacocke, C. 1984. Consciousness and other minds. *Aristotelian Society Supplement* 58:97-117.
- Plantinga, A. 1966. Induction and other minds. *Review of Metaphysics* 19:441-61.
- Plantinga, A. 1967. *God and Other Minds*. Cornell University Press.
- Plantinga, A. 1968. Induction and other minds II. *Review of Metaphysics* 12:524-33.
- Price, H. H. 1938. Our evidence for the existence of other minds. *Philosophy* 13:425-56.
- Sagal, P. & Borg, G. 1993. The range principle and the problem of other minds. *British Journal for the Philosophy of Science* 44:477-91.
- Slote, M. 1966. Induction and other minds. *Review of Metaphysics* 20:341-60.

- Sober, E. 2000. Evolution and the problem of other minds. *Journal of Philosophy* 97:365-387.
- Spencer, W. 1930. *Our Knowledge of Other Minds*. Yale University Press.
- Sprigge, T. L. S. 1992. Ayer on other minds. In (L. Hahn, ed) *The Philosophy of A. J. Ayer*. Open Court.
- Thalberg, I. 1969. Other times, other places, other minds. *Philosophical Studies* 20.
- Thomson, J. F. 1951. The argument from analogy and the problem of other minds. *Mind* 60:336-50.
- Weinberg, J. 1946. Our knowledge of other minds. *Philosophical Review* 60.
- Wisdom, J. 1946. Other minds. *Aristotelian Society Supplement* 20:122-47.
- Wisdom, J. 1968. *Other Minds*. University of California Press.
- Zemach, E. 1966. Sensations, raw feels, and other minds. *Review of Metaphysics* 20:317-40.
- Ziff, P. 1965. The simplicity of other minds. *Journal of Philosophy* 42:575-84.

Part 4: Philosophy of Artificial Intelligence [565]

Part of [Contemporary Philosophy of Mind: An Annotated Bibliography](#).

Compiled by [David J. Chalmers](#), Department of Philosophy, University of Arizona, Tucson AZ 85721. E-mail: chalmers@arizona.edu.

- 4.1 [Can Machines Think?](#) [224]
 - 4.1a [The Turing Test](#) [51]
 - 4.1b [Godelian Arguments](#) (Lucas, Penrose) [59]
 - 4.1c [The Chinese Room](#) (Searle) [58]
 - 4.1d [Machine Consciousness](#) [33]
 - 4.1e [Machine Thought, Misc](#) [23]
- 4.2 [Computation and Representation](#) [44]
 - 4.2a [Symbols and Symbol Systems](#) [9]
 - 4.2b [Computational Semantics](#) [16]
 - 4.2c [Implicit/Explicit Representation](#) [8]
 - 4.2d [AI without Representation?](#) [6]
 - 4.2e [Miscellaneous](#) [6]
- 4.3 [Philosophy of Connectionism](#) [150]
 - 4.3a [Connectionism and Compositionality](#) (Fodor/Pylyshyn) [43]
 - 4.3b [Representation in Connectionism](#) [20]
 - 4.3c [Connectionism and Eliminativism](#) [16]
 - 4.3d [The Connectionist/Classical Debate](#) [20]
 - 4.3e [Subsymbolic Computation](#) (Smolensky) [8]
 - 4.3f [Philosophy of Connectionism, Misc](#) [36]
 - 4.3g [Foundational Empirical Issues](#) [12]
- 4.4 [Dynamical Systems](#) [22]
- 4.5 [Foundational Questions in AI](#) [51]
 - 4.5a [The Nature of AI](#) [7]
 - 4.5b [Levels of Analysis](#) (Marr, etc) [12]
 - 4.5c [The Frame Problem](#) [14]
 - 4.5d [Methodology](#) [19]
- 4.6 [Computationalism in Cognitive Science](#) [33]
- 4.7 [Computation and Physical Systems](#) [18]
- 4.8 [Philosophy of AI, Misc](#) [17]

4.1 Can Machines Think?

4.1a

The Turing Test

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59:433-60.

Proposes the Imitation game (Turing test) as a test for intelligence: If a machine can't be told apart from a human in a conversation over a teletype, then that's good enough. With responses to various objections.

Alper, G. 1990. A psychoanalyst takes the Turing test. *Psychoanalytic Review* 77:59-68.

Barresi, J. 1987. Prospects for the Cyberiad: Certain limits on human self-knowledge in the cybernetic age. *Journal for the Theory of Social Behavior* 17:19-46.

Block, N. 1981. Psychologism and behaviorism. *Philosophical Review* 90:5-43.

A look-up table could pass the Turing test, and surely isn't intelligent. The TT errs in testing behavior and not mechanisms. A nice, thorough paper.

Bringsjord, S. 2001. Creativity, the Turing test, and the (better) Lovelace test. *Minds & Machines* 11:3-27.

Clark, T. 1992. The Turing test as a novel form of hermeneutics. *International Studies in Philosophy* 24:17-31.

Copeland, B. J. 2000. The Turing test. *Minds and Machines* 10:519-539.

Crawford, C. 1994. Notes on the Turing test. *Communications of the Association for Computing Machinery* 37:13-15.

Crockett, L. 1994. *The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence*. Ablex.

Davidson, D. 1990. Turing's test. In (K. Said, ed) *Modelling the Mind*. Oxford University Press.

Dennett, D. C. 1984. Can machines think? In (M. Shafto, ed) *How We Know*. Harper & Row.

Defending the Turing test as a good test for intelligence.

Erion, G. J. 2001. The Cartesian test for automatism. *Minds and Machines* 1:29-39.

French, R. M. 1990. Subcognition and the limits of the Turing test. *Mind* 99:53-66.

The Turing Test is too hard, as it requires not intelligence but human intelligence. Any machine could be unmasked through careful questioning, but this wouldn't mean that the machine was unintelligent.

French, R. M. 1995. Refocusing the debate on the Turing Test: A response. *Behavior and Philosophy* 23:59-60.

Response to Jacquette 1993.

Gunderson, K. 1964. The imitation game. *Mind* 73:234-45.

The Turing test is not broad enough: there's much more to thought than the ability to play the imitation game.

Harnad, S. 1991. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* 1:43-54.

On the Total Turing Test (full behavioral equivalence) as a test for mind.

Harnad, S. 1994. Levels of functional equivalence in reverse bioengineering: The Darwinian Turing test for artificial life. *Artificial Life* 1(3).

Harnad, S. 1999. Turing on reverse-engineering the mind. *Journal of Logic, Language, and Information*.

Hauser, L. 1993. Reaping the whirlwind: Reply to Harnad's "Other bodies, other minds". *Minds and Machines* 3:219-37.

Hauser, L. 2001. Look who's moving the goal posts now. *Minds and Machines* 11:41-51.

Hayes, P. & Ford, K. 1995. Turing test considered harmful. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 1:972-77.

Hofstadter, D. R. 1981. A coffee-house conversation on the Turing test. *Scientific American*.

A dialogue on the Turing test.

Jacquette, D. 1993. Who's afraid of the Turing test? *Behavior and Philosophy* 20:63-74.

Defending the Turing test against French 1990. Turing did not intend the test to provide a *necessary* condition for intelligence.

Jacquette, D. 1993. A Turing test conversation. *Philosophy* 68:231-33.

Karelis, C. 1986. Reflections on the Turing test. *Journal for the Theory of Social Behavior* 16:161-72.

Lee, E. T. 1996. On the Turing test for artificial intelligence. *Kybernetes* 25:61.

Leiber, J. 1989. Shanon on the Turing test. *Journal of Social Behavior*.

Leiber, J. 1995. On Turing's Turing Test and why the matter matters. *Synthese* 104:59-69.

Turing's test is neutral about the structure of the machine that passes it, but it must be practical and reliable (thus excluding Searle's and Block's counterexamples).

Mays, W. 1952. Can machines think? *Philosophy* 27:148-62.

Michie, D. 1993. Turing's test and conscious thought. *Artificial Intelligence* 60:1-22. Reprinted in (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Millar, P. 1973. On the point of the Imitation Game. *Mind* 82:595-97.

Moor, J. H. 1976. An analysis of Turing's test. *Philosophical Studies* 30:249-257.

The basis of the Turing test is not an operational definition of thinking, but rather an inference to the best explanation.

Moor, J. H. 1978. Explaining computer behavior. *Philosophical Studies* 34:325-7.

Reply to Stalker 1978: Mechanistic and mentalistic explanations are no more incompatible than program-based and physical explanations.

Moor, J. H. 2001. The status and future of the Turing test. *Minds and Machines* 11:77-93.

Piccinini, G. 2000. Turing's rules for the imitation game. *Minds and Machines* 10:573-582.

Purthill, R. 1971. Beating the imitation game. *Mind* 80:290-94.

Rankin, T. L. 1987. The Turing paradigm: A critical assessment. *Dialogue* 29:50-55.

Some obscure remarks on lying, imitation, and the Turing test.

Richardson, R. C. 1982. Turing tests for intelligence: Ned Block's defense of psychologism. *Philosophical Studies* 41:421-6.

A weak argument against Block: input/output function doesn't guarantee a capacity to respond sensibly.

Rosenberg, J. 1982. Conversation and intelligence. In (B. de Gelder, ed) *Knowledge and Representation*. Routledge & Kegan Paul.

Sampson, G. 1973. In defence of Turing. *Mind* 82:592-94.

Saygin, A. P. , Cicekli, I. & Akman V. 2000. Turing test: 50 years later. *Minds and Machines* 10:463-518.

Schweizer, P. 1998. The truly total Turing Test. *Minds and Machines* 8:263-272.

Shanon, B. 1989. A simple comment regarding the Turing test. *Journal for the Theory of Social Behavior* 19:249-56.

The Turing test presupposes a representational/computational framework for cognition. Not all phenomena can be captured in teletype communication.

Shieber, S. M. 1994. Lessons from a restricted Turing test. *Communications of the Association for Computing Machinery* 37:70-82.

Stalker, D. F. 1978. Why machines can't think: A reply to James Moor. *Philosophical Studies* 34:317-20.

Contra Moor 1976: The best explanation of computer behavior is mechanistic, not mentalistic.

Sterrett, S. G. 2000. Turing's two tests for intelligence. *Minds and Machines* 10:541-559.

Stevenson, J. G. 1976. On the imitation game. *Philosophia* 6:131-33.

Traiger, S. 2000. Making the right identification in the Turing test. *Minds and Machines* 10:561-572.

Watt, S. 1996. Naive psychology and the inverted Turing test. *Psychology* 7(14).

Whitby, B. 1996. The Turing test: AI's biggest blind alley? In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Zdenek, S. 2001. Passing Loebner's Turing test: A case of conflicting discourse functions. *Minds & Machines* 11:53-76.

4.1b

Godelian arguments (Lucas, Penrose)

Benacerraf, P. 1967. God, the Devil, and Godel. *Monist* 51:9-32.

Discusses and sharpens Lucas's arguments. Argues that the real consequence is that if we are Turing machines, we can't know which.

Bowie, G. 1982. Lucas' number is finally up. *Journal of Philosophy Logic*, 11:279-85.

Lucas's very Godelization procedure makes him inconsistent, unless he has an independent way to see if any TM is consistent, which he doesn't.

Boyer, D. 1983. J. R. Lucas, Kurt Godel, and Fred Astaire. *Philosophical Quarterly* 33:147-59.

Remarks on the various ways in which Lucas and a machine might be said to "prove" anything, and the ways in which a machine might simulate Lucas. The argument has all sorts of level confusions, and a bit of circularity.

Chari, C. 1963. Further comments on minds, machines and Godel. *Philosophy* 38:175-8.

Can't reduce the lawless creative process to computation.

Chalmers, D. J. 1996. Minds, machines, and mathematics. *Psyche* 2:11-20.

Chihara, C. 1972. On alleged refutations of mechanism using Godel's incompleteness results. *Journal of Philosophy* 64:507-26.

An analysis of the Lucas/Benacerraf argument. On various senses in which a machine might come to know its own program.

Coder, D. 1969. Godel's theorem and mechanism. *Philosophy* 44:234-7.

Only mathematicians understand Godel, so Lucas's argument isn't general; and Turing machines can go wrong. Weak.

Dennett, D. C. 1978. The abilities of men and machines. In *Brainstorms*. MIT Press.

There is no unique TM which we are -- there could be many.

Edis, T. 1998. How Godel's theorem supports the possibility of machine intelligence. *Minds and Machines* 8:251-262.

Feferman, S. 1996. Penrose's Godelian argument. *Psyche* 2:21-32.

Gaifman, H. 2000. What Godel's incompleteness result does and does not show. *Journal of Philosophy* 97:462-471.

George, F. 1962. Minds, machines and Godel: Another reply to Mr. Lucas. *Philosophy* 37:62-63.

Lucas's argument applies only to deductive machines, not inductive ones.

George, A. & Velleman, D. J. 2000. Leveling the playing field between mind and machine: A reply to McCall. *Journal of Philosophy* 97:456-452.

Good, I. J. 1967. Human and machine logic. *British Journal for the Philosophy of Science* 18:145-6.

Even humans can't Godelize forever. On ordinals and transfinite counting.

Good, I. J. 1969. Godel's theorem is a red herring. *British Journal for the Philosophy of Science* 19:357-8.

Rejoinder to Lucas 1967: the role of consistency; non-constructible ordinals.

Grush, R. & Churchland, P. 1995. Gaps in Penrose's toiling. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Hanson, W. 1971. Mechanism and Godel's theorem. *British Journal for the Philosophy of Science* 22:9-16.

An analysis of Benacerraf 1967. Benacerraf's "paradox" is illusory; there are no strong consequences of Godel's theorem for mechanism.

Hofstadter, D. R. 1979. *Godel, Escher, Bach: An Eternal Golden Braid*. Basic Books.

Contra Lucas: we can't Godelize forever; and we're not formal on top level.

Hutton, A. 1976. This Godel is killing me. *Philosophia* 3:135-44.

Gives a statistical argument to the effect that we cannot know that we are consistent; so the Lucas argument cannot go through.

Irvine, A. D. 1983. Lucas, Lewis, and mechanism -- one more time. *Analysis* 43:94-98.

Contra Lewis 1979, Lucas can derive the consistency of M even without the premise that he is M. Hmm.

Hadley, R. F. 1987. Godel, Lucas, and mechanical models of mind. *Computational Intelligence* 3:57-63.

A nice analysis of Lucas's argument and the circumstances under which a machine might prove another's Godel sentences. There's no reason to believe that machines and humans are different here.

Jacquette, D. 1987. Metamathematical criteria for minds and machines. *Erkenntnis* 27:1-16.

A machine will fail a Turing test if it's asked about Godel sentences.

King, D. 1996. Is the human mind a Turing machine? *Synthese* 108:379-89.

Kirk, R. 1986. Mental machinery and Godel. *Synthese*.

Lucas's argument fails, as theorems by humans don't correspond to outputs of their formal systems.

Lewis, D. 1969. Lucas against mechanism. *Philosophy* 44:231-3.

Lucas needs a rule of inference from sentences to their consistency, yielding Lucas arithmetic. No machine can prove all of Lucas arithmetic, but there's no reason to suppose humans can either, as the rule is infinitary.

Lewis, D. 1979. Lucas against mechanism II. *Canadian Journal of Philosophy* 9:373-6.

Reply to Lucas 1970: the dialectical argument fails, as the human's output depends on the premise that it is the machine (to derive M's consistency). With a similar premise, the machine itself can do equally well.

Lucas, J. R. 1961. Minds, machines and Godel. *Philosophy* 36:112-127.

Humans can Godelize any given machine, so we're not a machine.

Lucas, J. R. 1967. Human and machine logic: a rejoinder. *British Journal for the Philosophy of Science* 19:155-6.

Reply to Good 1967: a human can trump any given machine, so the human is not the machine, whether or not the human is superior across the board.

Lucas, J. R. 1968. Satan stultified: A rejoinder to Paul Benacerraf. *Monist* 52:145-58.

Benacerraf 1967 is empty and omega-inconsistent. Reply to arguments based on difficulty of seeing consistency (e.g. Putnam). Fallacious but engaging.

Lucas, J. R. 1971. Metamathematics and the philosophy of mind: A rejoinder. *Philosophy of Science* 38:310-13.

Lucas, J. R. 1970. Mechanism: A rejoinder. *Philosophy* 45:149-51.

Response to Lewis 1969 and Coder 1969. Lewis misses the dialectical nature of the argument.

Lucas, J. R. 1970. *The Freedom of the Will*. Oxford University Press.

Lucas, J. R. 1976. This Godel is killing me: A rejoinder. *Philosophia* 6:145-8.

Contra Hutton, we know -- even if fallibly -- that we are consistent.

Lucas, J. R. 1984. Lucas against mechanism II: A rejoinder. *Canadian Journal of Philosophy* 14:189-91.

Reply to Lewis 1979.

Lucas, J. R. 1996. Mind, machines and Godel: A retrospect. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Addresses all the counterarguments. Fun.

Lyngzeidetsen, A. E. & Solomon, M. K. 1994. Abstract complexity theory and the mind-machine problem. *British Journal for the Philosophy of Science* 45:549-54.

Lyngzeidetsen, A. 1990. Massively parallel distributed processing and a computationalist foundation for cognitive science. *British Journal for the Philosophy of Science* 41.

A Connection Machine might escape the Lucas argument. Bizarre.

Martin, J. & Engleman, K. 1990. The mind's I has two eyes. *Philosophy* 510-16.

Contra Hofstadter: Lucas can believe his Whitely sentence.

Maudlin, T. 1996. Between the motion and the act... *Psyche* 2:40-51.

McCall, S. 1999. Can a Turing machine know that the Godel sentence is true? *Journal of Philosophy* 96:525-32.

McCullough, D. 1996. Can humans escape Godel? *Psyche* 2:57-65.

McDermott, D. 1996. [Star] Penrose is wrong. *Psyche* 2:66-82.

Penrose, R. 1989. *The Emperor's New Mind*. Oxford University Press.

We are non-algorithmic as we can see Godel sentences of any algorithm.

Penrose, R. 1990. *Precis of The Emperor's New Mind*. Behavioral and Brain Sciences 13:643-705.

Much debate over the "non-algorithmic insight" in seeing Godel sentences.

Penrose, R. 1992. Setting the scene: The claim and the issues. In (D. Broadbent, ed) *The Simulation of Human Intelligence*. Blackwell.

An argument from the halting problem to the nonalgorithmicity of mathematical thought. Addresses objections: that the algorithm is unknowable, unsound, everchanging, environmental, or random. New physical laws may be involved.

Penrose, R. 1994. *Shadows of the Mind*. Oxford University Press.

Penrose, R. 1996. Beyond the doubting of a shadow. *Psyche* 2:89-129.

A reply to Chalmers, Feferman, Maudlin, McDermott, etc.

Priest, G. 1994. Godel's theorem and the mind... again. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind: The Place of Philosophy in the Study of Mind*. Kluwer.

Putnam, H. 1985. Reflexive reflections. *Erkenntnis* 22:143-153.

A generalized Godelian argument: if our prescriptive inductive competence is formalizable, then we could not know that such a formalization is correct.

Robinson, W. S. 1992. Penrose and mathematical ability. *Analysis* 52:80-88.

Penrose's argument depends on our knowledge of the validity of the algorithm we use, and here he equivocates between conscious and unconscious algorithms.

Slezak, P. 1982. Godel's theorem and the mind. *British Journal for the Philosophy of Science* 33:41-52.

General analysis; Lucas commits type/token error; self-ref paradoxes.

Slezak, P. 1983. Descartes's diagonal deduction. *British Journal for the Philosophy of Science* 34:13-36.

Cogito was a diagonal argument; connection to Godel, Lucas, Minsky, Nagel.

Smart, J. J. C. 1961. Godel's theorem, Church's theorem, and mechanism. *Synthese* 13:105-10.

A machine could escape the Godelian argument by inductively ascertaining its own syntax. With comments on the relevance of ingenuity.

Tymoczko, T. 1991. Why I am not a Turing Machine: Godel's theorem and the philosophy of mind. In (J. Garfield, ed) *Foundations of Cognitive Science*. Paragon House.

Weak defense of Lucas; response to Putnam, Bowie, Dennett.

Wang, H. 1974. *From Mathematics to Philosophy*. London.

Webb, J. 1968. Metamathematics and the philosophy of mind. *Philosophy of Science* 35:156-78.

Webb, J. 1980. *Mechanism, Mentalism and Metamathematics*. Kluwer.

Whitely, C. 1962. Minds, machines and Godel: A reply to Mr. Lucas. *Philosophy* 37:61-62.

Humans get trapped too: "Lucas cannot consistently assert this formula".

Yu, Q. 1992. Consistency, mechanicalness, and the logic of the mind. *Synthese* 90:145-79.

4.1c The Chinese Room (Searle)

Searle, J. R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-57.

Implementing a program is not sufficient for mentality, as someone could e.g. implement a "Chinese-speaking" program without understanding Chinese. So strong AI is false, and no program is sufficient for consciousness.

Searle, J. R. 1984. *Minds, Brains and Science*. Harvard University Press.

Axiomatizes the argument: Syntax isn't sufficient for semantics, programs are syntactic, minds are semantic, so no program is sufficient for mind.

Searle, J. R. 1987. Minds and brains without programs. In (C. Blakemore, ed) *Mindwaves*. Blackwell.

More on the arguments against AI, e.g. the Chinese room and considerations about syntax and semantics. Mind is a high-level physical property of brain.

Searle, J. R. 1990. Is the brain's mind a computer program? *Scientific American* 262(1):26-31.

On the status of the Chinese Room argument, ten years on.

Anderson, D. 1987. Is the Chinese room the real thing? *Philosophy* 62:389-93.

Boden, M. 1988. Escaping from the Chinese Room. In *Computer Models of Mind*. Cambridge University

Press.

A procedural account of how computers might have understanding and semantics.

Ben-Yami, H. 1993. A note on the Chinese room. *Synthese* 95:169-72.

A fully functional Chinese room is impossible, as it (for instance) could not say what the time is.

Bynum, T. W. 1985. Artificial intelligence, biology, and intentional states. *Metaphilosophy* 16:355-77.

A chess-playing machine embodied as a robot could have intentional states. Reference requires input/output, computation, and context.

Cam, P. 1990. Searle on strong AI. *Australasian Journal of Philosophy* 68:103-8.

Criticizes Searle's "conclusion" that brains are needed for intentionality, notes that even a homunculus has intentional states. A misinterpretation.

Carleton, L. 1984. Programs, language understanding, and Searle. *Synthese* 59:219-30.

Arguing against Searle on a number of fronts, somewhat unconvincingly.

Chalmers, D. J. 1992. Subsymbolic computation and the Chinese Room. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

Gives an account of symbolic vs. subsymbolic computation, and argues that the latter is less vulnerable to the Chinese-room intuition, as representations there are not computational tokens.

Churchland, P. M. & Churchland, P. S. 1990. Could a machine think? *Scientific American* 262(1):32-37.

Artificial mentality is possible, not through classical AI but through brain-like AI. Argues the syntax/semantics point using an analogy with electromagnetism and luminance.

Cohen, L. J. 1986. What sorts of machines can understand the symbols they use? *Aristotelian Society Supplement* 60:81-96.

Cole, D. J. 1984. Thought and thought experiments. *Philosophical Studies* 45:431-44.

Lots of thought experiments like Searle's, against Searle. Searle's argument is like Leibniz's "mill" argument, with similar level confusions. Nice but patchy.

Cole, D. J. 1991. Artificial intelligence and personal identity. *Synthese* 88:399-417.

In the Chinese room, neither the person nor the system understands: a virtual person does. This person isn't the system, just as a normal person isn't a body. Follows from the "Kornese" room, which has two distinct understanders.

Cole, D. J. 1991. Artificial minds: Cam on Searle. *Australasian Journal of Philosophy* 69:329-33.

Cole, D. J. 1994. The causal powers of CPUs. In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

Copeland, B. J. 1993. The curious case of the Chinese gym. *Synthese* 95:173-86.

Advocates the systems reply, and criticizes Searle's "Chinese Gym" response to connectionism: Searle (like those he accuses) confuses a simulation with the thing being simulated. Nice.

Dennett, D. C. 1987. Fast thinking. In *The Intentional Stance*. MIT Press.

Argues with Searle on many points. A little weak.

Double, R. 1983. Searle, programs and functionalism. *Nature and System* 5:107-14.

The homunculus doesn't have access to the system's intentionality. The syntax/semantics relation is like the neurophysiology/mind relation.

Dyer, M. 1990. Intentionality and computationalism: minds, machines, Searle and Harnad. *Journal of Experimental and Theoretical Artificial Intelligence* 2:303-19.

Reply to Searle/Harnad: systems reply, level confusions, etc.

Dyer, M. 1990. Finding lost minds. *Journal of Experimental and Theoretical Artificial Intelligence* 2:329-39.

Reply to Harnad 1990: symbols, other minds, physically embodied algorithms.

Fields, C. 1984. Double on Searle's Chinese Room. *Nature and System* 6:51-54.

Double's argument implies that the brain isn't the basis of intentionality.

Fisher, J. 1988. The wrong stuff: Chinese rooms and the nature of understanding. *Philosophical Investigations* 11:279-99.

Fodor, J. A. 1991. Yin and Yang in the Chinese Room. In (D. Rosenthal, ed) *The Nature of Mind*. Oxford University Press.

The Chinese room isn't even implementing a Turing machine, because it doesn't use proximal causation.
With a reply by Searle.

Globus, G. 1991. Deconstructing the Chinese room. *Journal of Mind and Behavior* 12:377-91.

Gozzano, S. 1995. Consciousness and understanding in the Chinese room. *Informatica* 19:653-56.

Hanna, P. 1985. Causal powers and cognition. *Mind* 94:53-63.

Argues that Searle is confused, and underestimates computers. Weak.

Harnad, S. 1989. Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence* 1:5-25.

Non-symbolic function is necessary for mentality. Trying hard to work out a theory of why the Chinese Room shows what it does. Nice but wrong.

Harnad, S. 1990. Lost in the hermeneutical hall of mirrors. *Journal of Experimental and Theoretical Artificial Intelligence* 2:321-27.

Reply to Dyer 1990: on the differences between real and as-if intentionality.

Hauser, L. 1997. Searle's Chinese box: Debunking the Chinese room argument. *Minds and Machines* 7:199-226.

Hayes, P. , Harnad, S. , Perlis, D. & Block, N. 1992. Virtual symposium on virtual mind. *Minds and Machines* 2.

A discussion about the Chinese room, symbol grounding, and so on.

Hofstadter, D. R. 1981. Reflections on Searle. In (D. Hofstadter & D. Dennett, eds) *The Mind's I*, pp. 373-382. Basic Books.

Searle is committing a level confusion, and understates the complexity of the case. We can move from the CR to a brain (with a demon) by twiddling knobs, and the systems reply should work equally well in both cases.

Jacquette, D. 1989. Searle's intentionality thesis. *Synthese* 80:267-75.

Searle's view implies that intentional causation is not efficient causation.

Jacquette, D. 1989. Adventures in the Chinese Room. *Philosophy and Phenomenological Research* 49:605-23.

If we had microfunctional correspondence, the CR argument would fail. With points about the status of abstract/biological intentionality. A bit weak.

Searle, J. R. 1989. Reply to Jacquette. *Philosophy and Phenomenological Research* 49:701-8.

Jacquette misses the point of the argument. Also, biological and abstract intentionality are quite compatible.

Jacquette, D. 1990. Fear and loathing (and other intentional states) in Searle's Chinese Room. *Philosophical Psychology* 3:287-304.

Reply to Searle on CR, central control, biological intentionality & dualism.

Jahren, N. 1990. Can semantics be syntactic? *Synthese* 82:309-28.

Against Rapaport's Korean Room argument -- syntax isn't enough.

Korb, K. 1991. Searle's AI program. *Journal of Experimental and Theoretical Artificial Intelligence* 3:283-96.

The Chinese room doesn't succeed as an argument about semantics. At best it might succeed as an argument about consciousness.

Maloney, J. C. 1987. The right stuff. *Synthese* 70:349-72.

Defends Searle against all kinds of objections.

Melnyk, A. 1996. Searle's abstract argument against strong AI. *Synthese* 108:391-419.

Moor, J. H. 1988. The pseudorealization fallacy and the Chinese Room argument. In (J. Fetzer, ed) *Aspects of AI*. D. Reidel.

Computational systems must also meet performance criteria.

Newton, N. 1989. Machine understanding and the Chinese Room. *Philosophical Psychology* 2:207-15.

A program can possess intentionality, even if not consciousness.

Obermeier, K. K. 1983. Wittgenstein on language and artificial intelligence: The Chinese-room thought-experiment revisited. *Synthese* 56:339-50.

Pfeifer, K. 1992. Searle, strong AI, and two ways of sorting cucumbers. *Journal of Philosophical Research*

17:347-50.

Rapaport, W. 1984. Searle's experiments with thought. *Philosophy of Science* 53:271-9.

Comments on Cole, and some general material on syntax and semantics.

Rey, G. 1986. What's really going on in Searle's 'Chinese Room'. *Philosophical Studies* 50:169-85.

Recommends the systems reply, and a causal account of semantics. Discusses the relevance of wide and narrow notions of content, and the tension between Searle's positive and negative proposals.

Roberts, L. 1990. Searle's extension of the Chinese Room to connectionist machines. *Journal of Experimental and Theoretical Artificial Intelligence* 2:185-7.

In arguing against the relevance of the serial/parallel distinction to mental states, Searle becomes a formalist. A nice point.

Russow, L-M. 1984. Unlocking the Chinese Room. *Nature and System* 6:221-8.

Searle's presence in the room destroys the integrity of the system, so that it is no longer a proper implementation of the program.

Seidel, A. 1988. Searle on the biological basis of cognition. *Analysis* 48:26-28.

Seidel, A. 1989. Chinese Rooms A, B and C. *Pacific Philosophical Quarterly* 20:167-73.

A person running the program, with interpretations at hand, would understand. Point-missing.

Sharvy, R. 1985. Searle on programs and intentionality. *Canadian Journal of Philosophy Supplement* 11:39-54.

Argues against Searle, but misses the point for the most part.

Sloman, A. 1986. Did Searle attack Strong Strong AI or Weak Strong AI? In (Cohn & Thomas, eds) *Artificial Intelligence and its Applications*. Chichester.

Suits, D. 1989. Out of the Chinese Room. *Computing and Philosophy Newsletter* 4:1-7.

Story about homunculi within homunculi. Fun.

Teng, N. Y. 2000. A cognitive analysis of the Chinese room argument. *Philosophical Psychology* 13:313-24.

Thagard, P. 1986. The emergence of meaning: An escape from Searle's Chinese Room. *Behaviorism* 14:139-46.

Get semantics computationally via induction and functional roles.

Weiss, T. 1990. Closing the Chinese room. *Ratio* 3:165-81.

Searle-in-the-room isn't in a position to know about the system's first-person states. Intrinsic intentionality is an incoherent notion.

Whitmer, J. M. 1983. Intentionality, artificial intelligence, and the causal powers of the brain. *Auslegung* 10:194-210.

Defending Searle's position, with remarks on the "causal powers" argument.

4.1d Machine Consciousness, Misc [see also 1.8b]

Angel, L. 1989. *How to Build a Conscious Machine*. Westview Press.

Angel, L. 1994. Am I a computer? In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

Arrington, R. 1999. Machines, consciousness, and thought. *Idealistic Studies* 29:231-243.

Barnes, E. 1991. The causal history of computational activity: Maudlin and Olympia. *Journal of Philosophy* 88:304-16.

Response to Maudlin 1989. True computation needs active, not passive causation, so Maudlin's machine isn't really computing.

Birnbacher, D. 1995. Artificial consciousness. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Bringsjord, S. 1992. *What Robots Can and Can't Be*. Kluwer.

Bringsjord, S. 1994. Could, how could we tell if, and should -- androids have inner lives? In (K. M. Ford, C. Glymour, & P. Hayes, eds) *Android Epistemology*. MIT Press.

Caplain, G. 1995. Is consciousness a computational property? *Informatica* 19:615-19.

Coles, L. S. 1993. Engineering machine consciousness. *AI Expert* 8:34-41.

- D'Aquili, E. G. & Newberg, A. B. 1996. Consciousness and the machine. *Zygon* 31:235-52.
- Danto, A. 1960. On consciousness in machines. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.
- Dennett, D. C. 1994. The practical requirements for making a conscious robot. *Philosophical Transactions of the Royal Society A* 349:133-46.
- Dennett, D. C. 1995. Cog: Steps toward consciousness in robots. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.
- Glennan, S. S. 1995. Computationalism and the problem of other minds. *Philosophical Psychology* 8:375-88.
- Gunderson, K. 1968. Robots, consciousness and programmed behaviour. *British Journal for the Philosophy of Science* 19:109-22.
- Gunderson, K. 1969. Cybernetics and mind-body problems. *Inquiry* 12:406-19.
- Gunderson, K. 1971. *Mentality and Machines*. Doubleday.
- Hillis, D. 1998. Can a machine be conscious? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Kirk, R. 1986. Sentience, causation and some robots. *Australasian Journal of Philosophy* 64:308-21.
- One could model brain states with monadic states and appropriate connections. But surely that's not intelligent -- the causation has the wrong form. Nice.
- Kitamura, T. , Tahara, T. , & Asami, K. 2000. How can a robot have consciousness? *Advanced Robotics* 14:263-275.
- Lucas, J. R. 1994. A view of one's own (conscious machines). *Philosophical Transactions of the Royal Society, Series A* 349:147-52.
- Maudlin, T. 1989. Computation and consciousness. *Journal of Philosophy* 86:407-32.
- Computational state is not sufficient for consciousness, as it can be instantiated by a mostly inert object. A nice thought-experiment, raising questions about the relevance of counterfactuals to consciousness.
- McCarthy, J. 1996. Making robots conscious of their mental states. In (S. Muggleton, ed) *Machine Intelligence 15*. Oxford University Press.

McGinn, C. 1987. Could a machine be conscious? In (C. Blakemore & S. Greenfield, ed) *Mindwaves*. Blackwell. Reprinted in *The Problem of Consciousness* (Blackwell, 1980).

Of course, as we are machines. But what *sort* of machines are conscious, and in virtue of what properties? Remarks on artefacts, life, functionalism, and computationalism. So far, we don't know what makes the brain conscious.

Puccetti, R. 1967. On thinking machines and feeling machines. *British Journal for the Philosophy of Science* 18:39-51.

Machines can think but can't feel, so aren't persons.

Putnam, H. 1964. Robots: machines or artificially created life? *Journal of Philosophy* 61:668-91. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Various arguments and counter-arguments re machine consciousness and civil liberties. Problems of machine consciousness are analogous to problems of human consciousness. The structural basis of the two may well be the same.

Putnam, H. 1967. The mental life of some machines. In (H. Castaneda, ed) *Intentionality, Minds and Perception*. Wayne State University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

More on TMs: explaining their psychology via preference functions.

Schlagel, R. 1999. Why not artificial consciousness or thought? *Minds and Machines* 9:3-28.

Scriven, M. 1953. The mechanical concept of mind. *Mind*.

To speak of a conscious machine is to commit a semantic mistake. Consciousness presupposes life and non-mechanism. Later retracted.

Stubenbergl, L. 1992. What is it like to be Oscar? *Synthese* 90:1-26.

Argues that AI systems like Pollock's Oscar needn't be conscious. Blindsight tells us that complex perceptual processing can go on unconsciously.

Thompson, D. 1965. Can a machine be conscious? *British Journal for the Philosophy of Science* 16:36.

Accepting machine consciousness would have few philosophical consequences, whereas rejecting it would tend to commit one to epiphenomenalism.

van de Vete, D. 1971. The problem of robot consciousness. *Philosophy and Phenomenological Research*

32:149-65.

Ziff, P. 1959. The feelings of robots. *Analysis*.

Of course robots can't think: they're not alive, so this gives us good reason not to rely on behavior. With replies by J.J.C. Smart, N. Smart.

4.1e Machine Thought, Misc

Bringsjord, S. 1998. Cognition is not computation: The argument from irreversibility. *Synthese*.

Burks, A. W. 1973. Logic, computers, and men. *Proceedings and Addresses of the American Philosophical Association* 46:39-57.

Arguing that a finite deterministic automaton can perform all natural human functions. With remarks on the logical organization of computers.

Cohen, L. J. 1955. Can there be artificial minds? *Analysis* 16:36-41.

Subservience to known or knowable rules is incompatible with mentality.

Copeland, B. J. 2000. Narrow versus wide mechanism: Including a re-examination of Turing's views on the mind-machine issue. *Journal of Philosophy* 97:5-33.

Dennett, D. C. 1985. Can machines think? In *How We Know* (Shafto).

Defends the Turing Test, among other things.

Dretske, F. 1985. Machines and the mental. *Proceedings and Addresses of the American Philosophical Association* 59:23-33.

Dretske, F. 1993. Can intelligence be artificial? *Philosophical Studies* 71:201-16.

Intelligence requires not just action or thought, but the governance of action by thought, which requires a history. "Wired-up" systems lack the explanatory connection between thought and action, so are not intelligent.

Dreyfus, H. L. 1972. *What Computers Can't Do*. Harper and Row.

Computers follow rules, people don't.

Hauser, L. 1993. Why isn't my pocket calculator a thinking thing? *Minds and Machines* 3:3-10.

Henley, T. B. 1990. Natural problems and artificial intelligence. *Behavior and Philosophy* 18:43-55.

On the philosophical importance of criteria for intelligence. With remarks on Searle, the Turing test, attitudes to AI, and ethical considerations.

Kearns, J. T. 1997. Thinking machines: Some fundamental confusions. *Minds and Machines* 7:269-87.

Lanier, J. 1998. Three objections to the idea of artificial intelligence. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Mackay, D. M. 1951. Mind-life behavior in artifacts. *British Journal for the Philosophy of Science* 2:105-21.

Mackay, D. M. 1952. Mentality in machines. *Aristotelian Society Supplement* 26:61-86.

Manning, R. C. 1987. Why Sherlock Holmes can't be replaced by an expert system. *Philosophical Studies* 51:19-28.

An expert system would lack Holmes' ability to raise the right questions, sort out relevant data, and determine what data are in need of explanation.

Mays, W. 1952. Can machines think? *Philosophy* 27:148-62.

McCarthy, J. 1979. Ascribing mental qualities to machines. In (M. Ringle, ed) *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

Negley, G. 1951. Cybernetics and theories of mind. *Journal of Philosophy* 48:574-82.

Preston, B. 1995. The ontological argument against the mind-machine hypothesis. *Philosophical Studies* 80:131-57.

Lucas, Searle, and Penrose all fall prey to "dual-description" fallacies.

Puccetti, R. 1966. Can humans think? *Analysis*.

Rapaport, W. 1993. Because mere calculating isn't thinking: Comments on Hauser's "Why isn't my pocket calculator a thinking thing?". *Minds and machines* 3:11-20.

Scriven, M. 1960. The compleat robot: A prolegomena to androidology. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

A machine could possess every characteristic of human thought: e.g. freedom, creativity, learning,

understanding, perceiving, feeling.

Spilsbury, R. J. 1952. Mentality in machines. *Aristotelian Society Supplement* 26:27-60.

4.2

Computation and Representation

4.2a

Symbols and Symbol Systems [see also 2.1a, 4.3e]

Cummins, R. 1996. Why there is no symbol grounding problem? In *Representations, Targets, and Attitudes*. MIT Press.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42:335-346.

AI symbols are empty and meaningless. They need to be "grounded" in something, e.g. sensory projection.
Maybe connectionism can do the trick?

Harnad, S. 1992. Connecting object to symbol in modeling cognition. In (A. Clark & R. Lutz, eds) *Connectionism in Context*. Springer-Verlag.

On the limitations of symbol systems, and the potential for grounding symbols in sensory icons and categorical perception, e.g. with neural networks.

Kosslyn, S. M. & Hatfield, G. 1984. Representation without symbol systems. *Social Research* 51:1019-1045.

MacDorman, K. F. 1997. How to ground symbols adaptively. In (S. O'Neill, P. McKevitt, & E. MacAogain, eds) *Two Sciences of Mind*. John Benjamins.

Newell, A. 1980. Physical symbol systems. *Cognitive Science* 4:135-83.

Newell, A. & Simon, H. A. 1981. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery* 19:113-26. Reprinted in (J. Haugeland, ed) *Mind Design*. MIT Press.

On computer science, AI, & the Physical Symbol System Hypothesis.

Robinson, W. S. 1995. Brain symbols and computationalist explanation. *Minds and Machines* 5:25-44.

Sun, R. 2000. Symbol grounding: a new look at an old idea. *Philosophical Psychology* 13:149-172.

4.2b Computational Semantics

Fodor, J. A. 1978. Tom Swift and his procedural grandmother. *Cognition* 6:229-47. Reprinted in *RePresentations* (MIT Press, 1980).

Against procedural semantics; it's a rerun of verificationism.

Hadley, R. F. 1990. Truth conditions and procedural semantics. In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

Johnson-Laird, P. 1977. Procedural semantics. *Cognition* 5:189-214.

Johnson-Laird, P. 1978. What's wrong with Grandma's guide to procedural semantics: A reply to Jerry Fodor. *Cognition* 9:249-61.

McDermott, D. 1978. Tarskian semantics, or no notation without denotation. *Cognitive Science* 2:277-82.

On the virtues of denotational semantics for AI. Notation without denotation, as found in many AI systems, leads to castles in the air.

Perlis, D. 1991. Putting one's foot in one's head -- Part 1: Why. *Nous* 25:435-55.

Perlis, D. 1994. Putting one's foot in one's head -- Part 2: How. In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

Rapaport, W. J. 1988. Syntactic semantics: Foundations of computational natural language understanding. In (J. Fetzer, ed) *Aspects of AI*. Kluwer.

Rapaport, W. J. 1995. Understanding understanding: Syntactic semantics and computational cognition. *Philosophical Perspectives* 9:49-88.

Smith, B. 1988. On the semantics of clocks. In (J. Fetzer, ed) *Aspects of AI*. Kluwer.

Smith, B. 1987. The correspondence continuum. *CSLI*-87-71.

Wilks, Y. 1982. Some thoughts on procedural semantics. In (W. Lehnert, ed) *Strategies for Natural Language Processing*. Lawrence Erlbaum.

Wilks, Y. 1990. Form and content in semantics. *Synthese* 82:329-51.

Criticism of McDermott's views on semantics, logic and natural language.

Winograd, T. 1985. Moving the semantic fulcrum. *Linguistics and Philosophy* 8:91-104.

Woods, W. 1981. Procedural semantics as a theory of meaning. In (A. Joshi, B. Weber, & I. Sag) *Elements of Discourse Understanding*. Cambridge University Press.

Woods, W. 1986. Problems in procedural semantics. In (Z. Pylyshyn & W. Demopolous, eds) *Meaning and Cognitive Structure*. Ablex.

With commentaries by Haugeland, J. D. Fodor.

4.2c

Implicit/Explicit Rules and Representations

Clark, A. 1991. In defense of explicit rules. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Argues that we need explicit rules for flexibility, adaptability, and representational redescription. With remarks on eliminativism.

Cummins, R. 1986. Inexplicit information. In (M. Brand & R. Harnish, eds) *The Representation of Knowledge and Belief*. University of Arizona Press.

On various kinds of representation of knowledge or belief without explicit tokens: control-implicit, domain-implicit, and procedural information. The key distinction is representation vs. execution of a rule.

Davies, M. 1995. Two notions of implicit rules. *Philosophical Perspectives* 9:153-83.

Hadley, R. F. 1990. Connectionism, rule-following, and symbolic manipulation. *Proc AAAI*.

Some rules are learnt so quickly that representation must be explicit.

Hadley, R. F. 1993. Connectionism, explicit rules, and symbolic manipulation. *Minds and Machines* 3.

Hadley, R. F. 1995. The 'explicit-implicit' distinction. *Minds and Machines* 5:219-42.

Kirsh, D. 1990. When is information explicitly represented? In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

Skokowski, P. G. 1994. Can computers carry content "inexplicitly"? *Minds and Machines* 4:333-44.

Cummins' account of inexplicit information fails, as even "executed" rules must be represented in the system. With remarks on the Chinese room.

4.2d

AI without Representation?

Bechtel, W. 1996. Yet another revolution? Defusing the dynamical system theorists' attack on mental representations. Manuscript.

Brooks, R. 1991. Intelligence without representation. *Artificial Intelligence* 47:139-159.

We don't need explicit representation; the world can do the job instead. Use embodied, complete systems, starting simple and working incrementally.

Clark, A. and Toribio, J. 1994. Doing without representing. *Synthese* 101:401-31.

A discussion of anti-representationalism in situated robotics and the dynamic systems movement (Brooks, Beer, van Gelder). These arguments appeal to overly simple domains, and a modest notion of representation survives.

Keijzer, F. A. 1998. Doing without representations which specify what to do. *Philosophical Psychology* 11:269-302.

Kirsh, D. 1991. Today the earwig, tomorrow man? *Artificial Intelligence* 47:161-184.

van Gelder, T. 1995. What might cognition be if not computation? *Journal of Philosophy* 92:345-81.

Argues for a dynamic-systems conception of the mind that is non-computational and non-representational. Uses an analogy with the Watt steam governor to argue for a new kind of dynamic explanation.

4.2e

Miscellaneous

Chrisley, R. L. 1994. Taking embodiment seriously: Nonconceptual content and robotics. In (K. M. Ford, C. Glymour, & P. Hayes, eds) *Android Epistemology*. MIT Press.

Dietrich, E. 1988. Computers, intentionality, and the new dualism. Manuscript.

Dreyfus, H. L. 1979. A framework for misrepresenting knowledge. In (M. Ringle, ed) *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

On the problems with context-free symbolic representation.

Fields, C. 1994. Real machines and virtual intentionality: An experimentalist takes on the problem of representational content. In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

Haugeland, J. 1981. Semantic engines: An introduction to mind design. In (J. Haugeland, ed) *Mind Design*. MIT Press.

Robinson, W. S. 1995. Direct representation. *Philosophical Studies* 80:305-22.

On Searle's critique of computational explanation, contrasted with Gallistel's use thereof. The real issue is computation on indirect vs. direct representations; direct computationalism is an attractive view.

4.3

Philosophy of Connectionism

4.3a

Connectionism and Compositionality (Fodor/Pylyshyn)

Fodor, J. A. & Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture. *Cognition* 28:3-71.

Connectionist models can't explain cognitive systematicity and productivity, as their representations lack compositional structure. The allures of connectionism are illusory; it's best used as an implementation strategy.

Aizawa, K. 1997. Explaining systematicity. *Mind and Language* 12:115-36.

Aizawa, K. 1997. The role of the systematicity argument in classicism and connectionism. In (S. O'Neill, ed) *Two Sciences of Mind*. John Benjamins.

Aizawa, K. 1997. Exhibiting versus explaining systematicity: A reply to Hadley and Hayward. *Minds and Machines* 7:39-55.

Antony, M. 1991. Fodor and Pylyshyn on connectionism. *Minds and Machines* 1:321-41.

Fodor and Pylyshyn's argument is an invalid instance of inference to the best explanation, as there is much

to explain than systematicity. Connectionism and classicism may be compatible even without implementation, in any case.

Aydede, M. 1997. Language of thought: The connectionist contribution. *Minds and Machines* 7:57-101.

Butler, K. 1991. Towards a connectionist cognitive architecture. *Mind and Language* 6:252-72.

Connectionism can make do with unstructured representations, as long have they have the right causal relations between them.

Butler, K. 1993. Connectionism, classical cognitivism, and the relation between cognitive and implementational levels of analysis. *Philosophical Psychology* 6:321-33.

Contra Chalmers 1993, F&P's argument doesn't apply at the implementational level. Contra Chater and Oaksford 1990, connectionism can't be purely implementational, but some implementational details can be relevant.

Butler, K. 1993. On Clark on systematicity and connectionism. *British Journal for the Philosophy of Science* 44:37-44.

Argues against Clark on holism and the conceptual truth of systematicity.

Butler, K. 1995. Compositionality in cognitive models: The real issue. *Philosophical Studies* 78:153-62.

Chalmers, D. J. 1990. Syntactic transformations on distributed representations. *Connection Science* 2:53-62.

An experimental demonstration that connectionist models can handle structure-sensitive operations in a non-classical way, transforming structured representations of active sentences to passive sentences.

Chalmers, D. J. 1993. Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology* 6:305-319.

Points out a structural flaw in F&P's argument, and traces the problem to a lack of appreciation of distributed representation. With some empirical results on structure sensitive processing, and some remarks on explanation.

Chater, N. & Oaksford, M. 1990. Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition* 34:93-107.

Implementation can make a difference at the algorithmic level.

Christiansen, M. H. & Chater, N. 1994. Generalization and connectionist language learning. *Mind and*

Language 9:273-87.

Cummins, R. 1996. Systematicity. *Journal of Philosophy* 93:591-614.

Fetzer, J. H. 1992. Connectionism and cognition: Why Fodor and Pylyshyn are wrong. In (A. Clark & R. Lutz, eds) *Connectionism in Context*. Springer-Verlag.

Fodor, J. A. & McLaughlin, B. P. 1990. Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition* 35:183-205.

Smolensky's weak compositionality is useless; and tensor product architecture can't support systematicity, as nonexistent tokens can't play a causal role.

Fodor, J. A. 1997. Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition* 62:109-19.

Garcia-Carpintero, M. 1996. Two spurious varieties of compositionality. *Minds and Machines* 6:159-72.

Garfield, J. 1997. Mentalese not spoken here: Computation, cognition, and causation. *Philosophical Psychology* 10:413-35.

Guarini, M. 1996. Tensor products and split-level architecture: Foundational issues in the classicism-connectionism debate. *Philosophy of Science* 63:S239-47.

Hadley, R. F. 1997. Cognition, systematicity, and nomic necessity. *Mind and Language* 12:137-53.

Hadley, R. F. 1994. Systematicity in connectionist language learning. *Mind and Language* 9:247-72.

Argues that existing connectionist models do not achieve an adequate systematicity in learning; they fail to generalize to handle structures with novel constituents.

Hadley, R. F. 1994. Systematicity revisited. *Mind and Language* 9:431-44.

Hadley, R. F. & Hayward, M. B. 1997. Strong semantic systematicity from Hebbian connectionist learning. *Minds and Machines* 7:1-55.

Hadley, R. F. 1997. Cognition, systematicity, and nomic necessity. *Mind and Language* 12:137-53.

Hadley, R. F. 1997. Explaining systematicity: A reply to Kenneth Aizawa. *Minds and Machines* 12:571-79.

Hawthorne, J. 1989. On the compatibility of connectionist and classical models. *Philosophical*

Psychology 2:5-16.

Localist connectionist models may not be able to handle structured presentation, but appropriate distributed models can.

Horgan, T. & Tienson, J. 1991. Structured representations in connectionist systems? In (Davis, ed) *Connectionism: Theory and Practice*.

A discussion of how connectionism might achieve "effective syntax" without implementing a classical system.

Matthews, R. J. 1994. Three-concept monte: Explanation, implementation, and systematicity. *Synthese* 101:347-63.

F&P deal a sucker bet: on their terms, connectionism could never give a a non-implementational explanation of systematicity, as the notions are construed in a manner specific to classical architectures.

Matthews, R. J. 1997. Can connectionists explain systematicity? *Mind and Language* 12:154-77.

McLaughlin, B. P. 1992. Systematicity, conceptual truth, and evolution. In *Philosophy and the Cognitive Sciences*.

Against responses to Fodor and Pylyshyn claiming that cognitive theories needn't explain systematicity. Contra Clark, the conceptual truth of systematicity won't help. Contra others, nor will evolution.

McLaughlin, B. P. 1993. The connectionism/classicism battle to win souls. *Philosophical Studies* 71.

Argues that no connectionist model so far has come close to explaining systematicity. Considers the models of Elman, Chalmers, and Smolensky.

Niklasson, L. F. & van Gelder, T. 1994. On being systematically connectionist. *Mind and Language* 9:288-302.

Pollack, J. B. 1990. Recursive distributed representations. *Artificial Intelligence* 46:77-105.

Develops a connectionist architecture -- recursive auto-associative memory -- that can recursively represent compositional structures in distributed form.

Rowlands, M. 1994. Connectionism and the language of thought. *British Journal for the Philosophy of Science* 45:485-503.

F&P's argument confuses constituent structure with logical/sentential structure. Connectionism is a psychotechnic project, whereas propositional description is a psychosemantic project.

Schroder, J. 1998. Knowledge of rules, causal systematicity, and the language of thought. *Synthese* 117:313-330.

Smolensky, P. 1987. The constituent structure of connectionist mental states. *Southern Journal of Philosophy Supplement* 26:137-60.

F&P ignore distributed representation and interaction effects.

Smolensky, P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159-216.

Develops a connectionist architecture that represents compositional structures as tensor products of distributed representations.

Smolensky, P. 1991. Connectionism, constituency and the language of thought. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Connectionism can do compositionality its own way, including both weak compositionality (with context effects) or strong compositionality (via tensor products).

Smolensky, P. 1995. Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In (C. Macdonald, ed) *Connectionism: Debates on Psychological Explanation*. Blackwell.

van Gelder, T. 1990. Compositionality: A connectionist variation on a classical theme. *Cognitive Science* 14:355-84.

Connectionism can do compositionality functionally. All one needs is the right functional relation between representations; physical concatenation is not necessary.

van Gelder, T. 1991. Classical questions, radical answers. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

On connectionism as a Kuhnian paradigm shift in cognitive science, with emphasis on the implications of functional compositionality and distributed representations.

4.3b Representation in Connectionism

Butler, K. 1995. Representation and computation in a deflationary assessment of connectionist cognitive science. *Synthese* 104:71-97.

Clark, A. 1989. Connectionism, nonconceptual content, and representational redescription. Manuscript.

On some troubles connectionism has with higher-order knowledge. Contrasts Cussins, Karmiloff-Smith on development. Subsymbols without symbols are blind.

Clark, A. 1993. *Associative Engines: Connectionism, Concepts, and Representational Change*. MIT Press.

Clark, A. & Karmiloff-Smith, A. 1994. The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language* 8:487-519.

On the importance of representational redescription, and on the limits of connectionist networks in cross-domain knowledge transfer. What does a true believer need, above behavior: conceptual combination, real-world fluency?

Cummins, R. 1991. The role of representation in connectionist explanation of cognitive capacities. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Connectionism isn't really radical. There's no new concept of representation or of learning, and cognition can still be the manipulation of semantically structured representations.

Cussins, A. 1990. The connectionist construction of concepts. In (M. Boden, ed) *The Philosophy of AI*. Oxford University Press.

Connectionism builds up concepts from the nonconceptual level. From nonconceptual content (e.g. perceptual experiences) to the emergence of objectivity.

Garzon, F. 2000. A connectionist defence of the inscrutability thesis. *Mind and Language* 15:465-480.

Garzón, F. 2000. State space semantics and conceptual similarity: reply to Churchland. *Philosophical Psychology* 13:77-96.

Goschke, T. & Koppelberg, D. 1990. Connectionism and the semantic content of internal representation. *Review of International Philosophy* 44:87-103.

Goschke, T. & Koppelberg, D. 1991. The concept of representation and the representation of concepts in connectionist models. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

On correlational semantics and context-dependent representations.

Hatfield, G. 1991. Representation and rule-instantiation in connectionist systems. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Some remarks on psychology & physiology. Even connectionism uses psychological concepts.

Hatfield, G. 1991. Representation in perception and cognition: Connectionist affordances. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Haybron, D. M. 2000. The causal and explanatory role of information stored in connectionist networks. *Minds and Machines* 10:361-380.

Laakso, A. & Cottrell, G. 2000. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical Psychology* 13:47-76.

Place, U. T. 1989. Toward a connectionist version of the causal theory of reference. *Acta Analytica* 4:71-97.

Ramsey, W. 1995. Rethinking distributed representation. *Acta Analytica* 10:9-25.

Ramsey, W. 1997. Do connectionist representations earn their explanatory keep? *Mind and Language* 12:34-66.

Argues that talk of representations has no explanatory role in connectionist theory, and can be discarded. It can't be understood along the lines of the teleo-informational or classical frameworks.

Schopman, J. & Shawky, A. 1996. Remarks on the impact of connectionism on our thinking about concepts. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Tye, M. 1987. Representation in pictorialism and connectionism. *Southern Journal of Philosophy Supplement* 26:163-184.

Pictorialism isn't compatible with language of thought, but connectionism might be.

van Gelder, T. 1991. What is the D in PDP? In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Argues that distributed representation is best analyzed in terms of superposition of representation, not in terms of extendedness.

4.3c

Connectionism and Eliminativism

Ramsey, W. , Stich, S. P. & Garon, J. 1991. Connectionism, eliminativism and the future of folk psychology. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Connectionism implies eliminativism, as connectionist systems do not have functionally discrete contentful states, and folk psychology is committed to functional discreteness of propositional attitudes.

Bickle, J. 1993. Connectionism, eliminativism, and the semantic view of theories. *Erkenntnis*.

Outlines the semantic view of scientific theories, and applies it to the connectionism/eliminativism debate. There's no reason why folk psychology shouldn't be reducible, in a homogeneous or heterogeneous way.

Botterill, G. 1994. Beliefs, functionally discrete states, and connectionist networks. *British Journal for the Philosophy of Science* 45:899-906.

Distinguishes active from dispositional beliefs: the former are realized discretely in activation patterns, the latter nondiscretely in weights, which is all that folk psychology needs.

Clapin, H. 1991. Connectionism isn't magic. *Minds and Machines* 1:167-84.

Commentary on Ramsey/Stich/Garon. Connectionism has symbols that interact, and has propositional modularity in processing if not in storage.

Clark, A. 1989. Beyond eliminativism. *Mind and Language* 4:251-79.

Connectionism needn't imply eliminativism, as higher levels may have a causal role, if not causal completeness. Also, it may not tell the whole story.

Clark, A. 1990. Connectionist minds. *Proceedings of the Aristotelian Society* 90:83-102.

Responding to eliminativist challenge via cluster analysis and recurrence.

Davies, M. 1991. Concepts, connectionism, and the language of thought. (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Argues that our conception of thought requires causal systematicity, which requires a language of thought. Connectionist systems are not causally systematic, so connectionism leads to eliminativism.

Egan, F. 1995. Folk psychology and cognitive architecture. *Philosophy of Science* 62:179-96.

Forster, M. & Saidel, E. 1994. Connectionism and the fate of folk psychology. *Philosophical Psychology* 7:437-52.

Contra Ramsey, Stich, and Garon, connectionist representations can be seen to be functionally discrete on an appropriate analysis of causal relevance.

Horgan, T. , and Tienson, J. 1995. Connectionism and the commitments of folk psychology. *Philosophical Perspectives* 9:127-52.

O'Brien, G. 1991. Is connectionism commonsense? *Philosophical Psychology* 4:165-78.

O'Leary-Hawthorne, J. 1994. On the threat of eliminativism. *Philosophical Studies* 74:325-46.

A dispositional construal of beliefs and desires can distinguish the relevant active states (via counterfactuals) and is compatible with FP, so internals can't threaten FP. With remarks on Davidson, overdetermination, etc.

Place, U. T. 1992. Eliminative connectionism: Its implications for a return to an empiricist/behaviorist linguistics. *Behavior and Philosophy* 20:21-35.

Ramsey, W. 1994. Distributed representation and causal modularity: A rejoinder to Forster and Saidel. *Philosophical Psychology* 7:453-61.

Upon examination, the model of Forster and Saidel 1994 does not exhibit features that are both distributed and causally discrete.

Smolensky, P. 1995. On the projectable predicates of connectionist psychology: A case for belief. In (C. Macdonald, ed) *Connectionism: Debates on Psychological Explanation*. Blackwell.

Stich, S. & Warfield, T. 1995. Reply to Clark and Smolensky: Do connectionist minds have beliefs? In (C. Macdonald, ed) *Connectionism: Debates on Psychological Explanation*. Blackwell.

4.3d

The Connectionist/Classical Debate

Adams, F. , Aizawa, K. & Fuller, G. 1992. Rules in programming languages and networks. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

The distinction between programming languages and networks is neutral on rule-following, etc, so there's nothing really new about connectionism.

Aizawa, K. 1994. Representations without rules, connectionism, and the syntactic argument. *Synthese* 101:465-92.

Bringsjord, S. 1991. Is the connectionist-logician debate one of AI's wonderful red herrings? *Journal of Theoretical and Experimental Artificial Intelligence* 3:319-49.

A detailed analysis purporting to show that connectionism and "logicism" are compatible, as Turing machines can do everything a neural network can. Entertaining, but misunderstands subsymbolic processing.

Broadbent, D. 1985. A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General* 114:189-92.

Distributed models are at the implementational, not computational, level.

Chandrasekaran, B. , Goel, A. & Allemang, D. 1988. Connectionism and information-processing abstractions. *AI Magazine* 24-34.

Connectionism won't affect AI too much, as AI is concerned with the information-processing (task) level. With greater modularity, connectionism will look more like traditional AI.

Corbi, J. E. 1993. Classical and connectionist models: Levels of description. *Synthese* 95:141-68.

Dawson, M. R. W. , Medler, D. A. , & Berkeley, I. S. N. 1997. PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology* 10:25-40.

Dennett, D. C. 1986. The logical geography of computational approaches: A view from the east pole. In (M. Brand & R. Harnish, eds) *The Representation of Knowledge and Belief*. University of Arizona Press.

Drawing the battle-lines: High Church Computationalism at the "East Pole", New Connectionism, Zen Holism, etc, at various locations on the "West Coast". With remarks on connectionism, and on AI as thought-experimentation.

Dennett, D. C. 1991. Mother Nature versus the walking encyclopedia. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Reiterating the value of connectionism, especially biological plausibility.

Dinsmore, J. (ed) 1992. *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

Dyer, M. 1991. Connectionism versus symbolism in high-level cognition. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Garson, J. W. 1991. What connectionists cannot do: The threat to Classical AI. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Connectionism and classicism aren't necessarily incompatible on symbolic discreteness, causal role, functional discreteness, constituency, representation of rules.

Garson, J. W. 1994. No representations without rules: The prospects for a compromise between paradigms in cognitive science. *Mind and Language* 9:25-37.

Garson, J. W. 1994. Cognition without classical architecture. *Synthese* 100:291-306.

Horgan, T. & Tienson, J. 1987. Settling into a new paradigm. *Southern Journal of Philosophy Supplement* 26:97-113.

On connectionism, basketball, and representation without rules. Responses to the "syntactic" and "semantic" arguments against connectionism. Nice.

Horgan, T. & Tienson, J. 1989. Representation without rules. *Philosophical Perspectives* 17:147-74.

Cognition uses structured representations without high-level rules, and connectionism is better at accounting for this. With remarks on exceptions to psychological laws, and the crisis in traditional AI.

Horgan, T. & Tienson, J. 1994. Representations don't need rules: Reply to James Garson. *Mind and Language* 9:1-24.

McClelland, J. L. & Rumelhart, D. E. 1985. Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General* 114:193-7.

Response to Broadbent 1985: Distributed models are at the algorithmic level. Elucidating the low-level/high-level relation via various analogies.

McLaughlin, B. P. & Warfield, F. 1994. The allure of connectionism reexamined. *Synthese* 101:365-400.

Argues that symbolic systems such as decision trees are as good at learning and pattern recognition as connectionist networks, and it is just as plausible that they are implemented in the brain.

Rey, G. 1991. An explanatory budget for connectionism and eliminativism. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Challenges connectionism to explain things that the classical approach seems to handle better: the structure, systematicity, causal role, and grain of propositional attitudes, their rational relations, and conceptual stability.

4.3e

Subsymbolic Computation (Smolensky)

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1-23.

Connectionism offers a complete account at the subsymbolic level, rather than an approximate account at the symbolic level.

Berkeley, I. 2000. What the #\$%! is a subsymbol? Minds and machines 10:1-14.

Chalmers, D. J. 1992. Subsymbolic computation and the Chinese Room. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

Explicates the distinction between symbolic and subsymbolic computation, and argues that connectionism can better handle the emergence of semantics from syntax, due to the non-atomic nature of its representations.

Clark, A. 1993. Superpositional connectionism: A reply to Marinov. Minds and Machines 3:271-81.

Hofstadter, D. R. 1983. Artificial intelligence: Subcognition as computation. In (F. Machlup, ed) *The Study of Information: Interdisciplinary Messages*. Wiley. Reprinted as "Waking up from the Boolean dream" in *Metamagical Themas*. Basic Books.

AI needs statistical emergence. For real semantics, symbols must be decomposable, complex, autonomous -
- i.e. active.

Marinov, M. 1993. On the spuriousness of the symbolic/subsymbolic distinction. Minds and Machines 3:253-70.

Argues with Smolensky: symbolic systems such as decision trees have all the positive features of neural networks (flexibility, lack of brittleness), and can represent concepts as sets of subconcepts. With a reply by Clark.

Rosenberg, J. 1990. Treating connectionism properly: Reflections on Smolensky. Psychological Research 52:163.

Rejects Smolensky's PTC, as the proper interaction of the microscopic and macroscopic levels would take a "miracle".

Smolensky, P. 1987. Connectionist AI, symbolic AI, and the brain. AI Review 1:95-109.

On connectionist networks as subsymbolic dynamic systems.

4.3f Philosophy of Connectionism, Misc.

Bechtel, W. 1985. Are the new PDP models of cognition cognitivist or associationist? Behaviorism 13:53-61.

Bechtel, W. 1986. What happens to accounts of mind-brain relations if we forgo an architecture of rules and representations? *Philosophy of Science Association* 1986, 159-71.

On the relationship between connectionism, symbol processing, psychology and neuroscience.

Bechtel, W. 1987. Connectionism and the philosophy of mind. *Southern Journal of Philosophy Supplement* 26:17-41. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Lots of questions about connectionism.

Bechtel, W. 1988. *Connectionism and rules and representation systems: Are they compatible?* *Philosophical Psychology* 1:5-16.

There's room for both styles within a single mind. The rule-based level needn't be autonomous; the connectionist level plays a role in pattern recognition, concepts, and so on.

Bechtel, W. & Abrahamson, A. 1990. *Beyond the exclusively propositional era.* *Synthese* 82:223-53.

An account of the shift from propositions to pattern recognition in the study of cognition: knowing-how, imagery, categorization, connectionism.

Bechtel, W. & Abrahamsen, A. A. 1992. *Connectionism and the future of folk psychology.* In (R. Burton, ed) *Minds: Natural and Artificial.* SUNY Press.

Bechtel, W. 1993. The case for connectionism. *Philosophical Studies* 71:119-54.

Bickle, J. 1995. Connectionism, reduction, and multiple realizability. *Behavior and Philosophy* 23:29-39.

Bradshaw, D. E. 1991. Connectionism and the specter of representationalism. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind.* Kluwer.

Argues that connectionism allows for a more plausible epistemology of perception, compatible with direct realism rather than representationalism. With remarks on Fodor and Pylyshyn's argument against Gibson.

Churchland, P. M. 1989. On the nature of theories: A neurocomputational perspective. *Minnesota Studies in the Philosophy of Science* 14. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Connectionism will revolutionize our review of scientific theories: >From the deductive-nomological view to descent in weight-space. Some cute analogies.

Churchland, P. M. 1989. On the nature of explanation: A PDP approach. In *A Neurocomputational Perspective.* MIT Press.

We achieve explanatory understanding not through the manipulation of propositions but through the activation of prototypes.

Churchland, P. S. & Sejnowski, T. 1989. Neural representation and neural computation. In (L. Nadel, ed) *Neural Connections, Mental Computations*. MIT Press.

Implications of connectionism and neuroscience for our concept of mind.

Clark, A. 1989. *Microcognition*. MIT Press.

All kinds of stuff on connectionism and philosophy.

Clark, A. 1990. Connectionism, competence and explanation. *British Journal for the Philosophy of Science* 41:195-222.

Connectionism separates processing from competence. Instead of hopping down Marr's levels (theory->process), connectionism goes (1) task (2) low-level performance (3) extract theory from process. Cute.

Cummins, R. & Schwarz, G. 1987. Radical connectionism. *Southern Journal of Philosophy Supplement* 26:43-61.

On computation and representation in AI and connectionism, and on problems for radical connectionism in reconciling these without denying representation or embracing mystery.

Cummins, R. & Schwarz, G. 1991. Connectionism, computation, and cognition. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Explicates computationalism, and discusses ways in which connectionism might end up non-computational: if causal states cross-classify representational states, or if transitions between representations aren't computable.

Cummins, R. 1995. Connectionist and the rationale constraint on cognitive explanations. *Philosophical Perspectives* 9:105-25.

Davies, M. 1989. Connectionism, modularity and tacit knowledge. *British Journal for the Philosophy of Science* 40:541-55.

Argues that connectionist networks don't have tacit knowledge of modular theories (as representations lack the appropriate structure, etc.).

Globus, G. G. 1992. Derrida and connectionism: Differance in neural nets. *Philosophical Psychology* 5:183-97.

Hatfield, G. 1990. Gibsonian representations and connectionist symbol-processing: prospects for unification. *Psychological Research* 52:243-52.

Gibson is compatible with connectionism. In both, we can have rule-instantiation without rule-following.

Horgan, T. & Tienson, J. (eds) 1991. *Connectionism and the Philosophy of Mind*. Kluwer.

Horgan, T. & Tienson, J. 1996. *Connectionism and the Philosophy of Psychology*. MIT Press.

Horgan, T. 1997. Connectionism and the philosophical foundations of cognitive science. *Metaphilosophy* 28:1-30.

Humphreys, G. W. 1986. Information-processing systems which embody computational rules: The connectionist approach. *Mind and Language* 1:201-12.

Legg, C. R. 1988. Connectionism and physiological psychology: A marriage made in heaven? *Philosophical Psychology* 1:263-78.

Litch, M. 1997. Computation, connectionism and modelling the mind. *Philosophical Psychology* 10:357-364.

Lloyd, D. 1989. Parallel distributed processing and cognition: Only connect? In *Simple Minds*. MIT Press.

An overview: local/distributed/featural representations; explanation in connectionism (how to avoid big mush); relation to neuroscience; explicit representations of rules vs weight matrices.

Lycan, W. G. 1991. Homuncular functionalism meets PDP. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

On various ways in which connectionism relates to representational homuncular functionalism, e.g. on implementation, eliminativism, and explanation.

Macdonald, C. 1995. *Connectionism: Debates on Psychological Explanation*. Blackwell.

Ramsey, W. & Stich, S. P. 1990. Connectionism and three levels of nativism. *Synthese* 82:177-205.

How connectionism bears on the nativism debate. Conclusion: not too much.

Ramsey, W. , Stich, S. P. & Rumelhart, D. M. (eds) 1991. *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Rosenberg, J. 1989. Connectionism and cognition. Bielefeld Report.

Criticism of Churchland's connectionist epistemology.

Sehon, S. 1998. Connectionism and the causal theory of action explanation. *Philosophical Psychology* 11:511-532.

Shanon, B. 1992. Are connectionist models cognitive? *Philosophical Psychology*.

In some senses of "cognitive", yes; in other senses, no. Phenomenological, theoretical, and sociological perspectives. Toward meaning-laden models.

Sterelny, K. 1990. Connectionism. In *The Representational Theory of Mind*. Blackwell.

Waskan, J. & Bechtel, W. 1997. Directions in connectionist research: Tractable computations without syntactically structured representations. *Metaphilosophy* 28:31-62.

4.3g

Foundational Empirical Issues

Clark, A. 1994. Representational trajectories in connectionist learning. *Minds and Machines* 4:317-32.

On how to get connectionist networks to learn about structured task domains. Concentrates on incremental learning, and other developmental/scaffolding strategies. With remarks on systematicity.

Clark, A. & Thornton, S. 1997. Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences* 20:57-66.

Cliff, D. 1990. Computational neuroethology: A provisional manifesto. Manuscript.

Criticizes connectionism for not being sufficiently rooted in neuroscience, and for not being grounded in the world.

Dawson, M. R. W. & Schopflocher, D. P. 1992. Autonomous processing in parallel distributed processing networks. *Philosophical Psychology* 5:199-219.

Hanson, S. & Burr, D. 1990. What connectionist models learn. *Behavioral and Brain Sciences*.

What's new to connectionism is not learning or representation but the way that learning and representation interact.

Kaplan, S. , Weaver, M. & French, R. M. 1990. Active symbols and internal models: Towards a cognitive

connectionism. AI and Society.

Addresses behaviorist/associationist charges. Connectionism needs recurrent circuits to support active symbols.

Kirsh, D. 1987. Putting a price on cognition. *Southern Journal of Philosophy Supplement* 26:119-35.

On the importance of variable binding, and why it's hard with connectionism.

Lachter, J. & Bever, T. 1988. The relation between linguistic structure and associative theories of language learning. *Cognition* 28:195-247.

Criticism of connectionist language models. They build in too much.

Mills, S. 1989. Connectionism, the classical theory of cognition, and the hundred step constraint. *Acta Analytica* 4:5-38.

Nelson, R. 1989. Philosophical issues in Edelman's neural darwinism. *Journal of Experimental and Theoretical Artificial Intelligence* 1:195-208.

On the relationship between ND, PDP and AI. All are computational.

Oaksford, M. , Chater, N. & Stenning, K. 1990. Connectionism, classical cognitive science and experimental psychology. *AI and Society*.

Connectionism is better at explaining empirical findings about mind.

Pinker, S. & Prince, A. 1988. On language and connectionism. *Cognition* 28:73-193.

Extremely thorough criticism of the R&M past-tense-learning model, with arguments on why connectionism can't handle linguistic rules.

4.4

Dynamical Systems

Bechtel, W. 1996. Yet another revolution? Defusing the dynamical system theorists' attack on mental representations. Manuscript.

Clark, A. 1998. Time and mind. *Journal of Philosophy* 95:354-76.

Eliasmith, C. 1996. The third contender: A critical examination of the dynamicist theory of cognition.

Philosophical Psychology 9:441-63.

Eliasmith, C. 1997. Computation and dynamical models of mind. *Minds and Machines* 7:531-41.

Foss, J. E. 1992. Introduction to the epistemology of the brain: Indeterminacy, micro-specificity, chaos, and openness. *Topoi* 11:45-57.

On the brain as a vector-processing system, and the problems raised by indeterminacy, chaos, and so on.
With morals for cognitive science.

Freeman, W. 1997. Nonlinear neurodynamics of intentionality. *Journal of Mind and Behavior* 18:291-304.

Garson, J. W. 1995. Chaos and free will. *Philosophical Psychology* 8:365-74.

Garson, J. W. 1996. Cognition poised at the edge of chaos: A complex alternative to a symbolic mind. *Philosophical Psychology* 9:301-22.

Garson, J. W. 1997. Syntax in a dynamic brain. *Synthese* 110:343-355.

Garson, J. W. 1998. Chaotic emergence and the language of thought. *Philosophical Psychology* 11:303-315.

Giunti, M. 1995. Dynamic models of cognition. In (T. van Gelder & R. Port, eds) *Mind as Motion*. MIT Press.

Giunti, M. 1996. *Computers, Dynamical Systems, and the Mind*. Oxford University Press.

Globus, G. 1992. Toward a noncomputational cognitive science. *Journal of Cognitive Neuroscience* 4:299-310.

Hooker, C. A. & Christensen, W. D. 1998. Towards a new science of the mind: Wide content and the metaphysics of organizational properties in nonlinear dynamic models. *Mind and Language* 13:98-109.

Horgan, T. & Tienson, J. 1992. Cognitive systems as dynamic systems. *Topoi* 11:27-43.

Horgan, T. & Tienson, J. 1994. A nonclassical framework for cognitive science. *Synthese* 101:305-45.

Keijzer, F. A. & Bem, S. 1996. Behavioral systems interpreted as autonomous agents and as coupled dynamical systems: A criticism. *Philosophical Psychology* 9:323-46.

Sloman, A. 1993. The mind as a control system. In (C. Hookway & D. Peterson, eds) *Philosophy and*

Cognitive Science. Cambridge University Press.

van Gelder, T. & Port, R. 1995. *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press.

van Gelder, T. 1995. What might cognition be if not computation? *Journal of Philosophy* 92:345-81.

Argues for a dynamic-systems conception of the mind that is non-computational and non-representational.
Uses an analogy with the Watt steam governor to argue for a new kind of dynamic explanation.

van Gelder, T. 1997. Connectionism, dynamics, and the philosophy of mind. In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.

van Gelder, T. 1998. The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* 21:615-28.

4.5

Foundational Questions in AI

4.5a

The Nature of AI

Buchanan, B. 1988. AI as an experimental science. In (J. Fetzer, ed) *Aspects of AI*. D. Reidel.

Bundy, A. 1990. What kind of field is AI? In (D. Partridge & Y. Wilks, eds) *The Foundations of Artificial Intelligence: A Sourcebook*. Cambridge University Press.

Dennett, D. C. 1978. AI as philosophy and as psychology. In (M. Ringle, ed) *Philosophical Perspectives on Artificial Intelligence*. Humanities Press. Reprinted in *Brainstorms* (MIT Press, 1978).

AI as detailed armchair psychology and as thought-experimental epistemology. Implications for mind: e.g. a solution to the problem of homuncular regress.

Glymour, C. 1988. AI is philosophy. In (J. Fetzer, ed) *Aspects of AI*. D. Reidel.

Kukla, A. 1989. Is AI an empirical science? *Analysis* 49:56-60.

No, AI is an a priori science that uses empirical methods; its status is similar to that of mathematics.

Kukla, A. 1994. Medium AI and experimental science. *Philosophical Psychology* 7:493-5012.

On the status of "medium AI", the study of intelligence in computational systems (not just humans). Contra to many, this is not an empirical science, but a combination of (experimental) mathematics and engineering.

Nakashima, H. 1999. AI as complex information processing. *Minds and Machines* 9:57-80.

4.5b

Levels of Analysis (Marr, etc)

Bechtel, W. 1994. Levels of description and explanation in cognitive science. *Minds and Machines* 4:1-25.

Cleeremans, A. & French, R. M. 1996. From chicken squawking to cognition: Levels of description and the computational approach in psychology. *Psychologica Belgica* 36:5-29.

Foster, C. 1990. *Algorithms, abstraction and implementation*. Academic Press.

Outlines a theory of the equivalence of algorithms.

Horgan, T. & Tienson, J. 1992. Levels of description in nonclassical cognitive science. *Philosophy* 34, Supplement.

Generalizes Marr's levels to: cognitive state-transitions, mathematical state-transitions, implementation. Discusses these with respect to connectionism, dynamical systems, and computation below the cognitive level.

Houng, Y. 1990. Classicism, connectionism and the concept of level. Dissertation, Indiana University.

On levels of organization vs. levels of analysis.

Marr, D. 1982. *Vision*. Freeman.

Defines computational, algorithmic and implementational levels.

McClamrock, R. 1990. Marr's three levels: a re-evaluation. *Minds and Machines* 1:185-196.

On different kinds of level-shifts: organizational and contextual changes. There are more than three levels available.

Newell, A. 1982. The knowledge level. *Artificial Intelligence* 18:81-132.

Newell, A. 1986. The symbol level and the knowledge level. In (Z. Pylyshyn & W. Demopolous, eds) *Meaning and Cognitive Structure*. Ablex.

With commentaries by Smith, Dennett.

Peacocke, C. 1986. Explanation in computational psychology: Language, perception and level 1.5. *Mind and Language* 1:101-23.

Psychological explanation is typically somewhere *between* the computational and algorithmic levels.

Sticklen, J. 1989. Problem-solving architectures at the knowledge level. *Journal of Experimental and Theoretical Artificial Intelligence* 1:233-247.

4.5c The Frame Problem

Dennett, D. C. 1984. Cognitive wheels: The frame problem of AI. In (Hookaway, ed) *Minds, Machines and Evolution*. Cambridge University Press.

General overview.

Dreyfus, H. L. & Dreyfus, S. 1987. How to stop worrying about the frame problem even though it's computationally insoluble. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

FP is an artifact of computational explicitness. Contrast human commonsense know-how, with relevance built in. Comparison to expert/novice distinction.

Fetzer, J. H. 1990. The frame problem: Artificial intelligence meets David Hume. *International Journal of Expert Systems* 3:219-232.

Fodor, J. A. 1987. Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

FP is Hamlet's problem: when to stop thinking. It's equivalent to the general problem of non-demonstrative inference.

Haugeland, J. 1987. An overview of the frame problem. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

The FP may be a consequence of the explicit/implicit rep distinction. Use "complicit" reps instead, and changes will be carried along intrinsically.

Hayes, P. 1987. What the frame problem is and isn't. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

FP is a relatively narrow problem, Some, e.g. Fodor, wrongly equate FP with the "Generalized AI Problem".

Janlert, L. 1987. Modeling change: The frame problem. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

Korb, K. 1998. The frame problem: An AI fairy tale. *Minds and Machines* 8:317-351.

Lormand, E. 1990. Framing the frame problem. *Synthese* 82:353-74.

Criticizes Dennett's, Haugeland's and Fodor's construals of the FP.

Maloney, J. C. 1988. In praise of narrow minds. In (J. Fetzer, ed) *Aspects of AI*. D. Reidel.

McCarthy, J. & Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. In (Meltzer & Michie, eds) *Machine Intelligence 4*. Edinburgh University Press.

McDermott, D. 1987. We've been framed: Or, Why AI is innocent of the frame problem. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

Solve frame problem by using the sleeping-dog strategy -- keeping things fixed unless there's a reason to suppose otherwise.

Pollock, JL. 1997. Reasoning about change and persistence: A solution to the frame problem. *Nous* 31:143-169.

Pylyshyn, Z. W. (ed) 1987. *The Robot's Dilemma*. Ablex.

Lots of papers on the frame problem.

4.5d

AI Methodology

Birnbaum, L. 1991. Rigor mortis: A response to Nilsson's 'Logic and artificial intelligence'. *Artificial Intelligence* 47:57-78.

Chalmers, D. J. , French, R. M. & Hofstadter, D. R. 1992. High-level perception, representation, and analogy: A critique of AI methodology. *Journal of Experimental and Theoretical Artificial Intelligence*.

AI must integrate perception and cognition in the interest of flexible representation. Current models ignore

perception and the development of representation, but this cannot be separated from later cognitive processes.

Clark, A. 1986. A biological metaphor. *Mind and Language* 1:45-64.

AI should look at biology.

Clark, A. 1987. The kludge in the machine. *Mind and Language* 2:277-300.

Dascal, M. 1992. Why does language matter to artificial intelligence? *Minds and Machines* 2:145-174.

Dreyfus, H. L. 1981. From micro-worlds to knowledge: AI at an impasse. In (J. Haugeland, ed) *Mind Design*. MIT Press.

Micro-worlds don't test true understanding, and frames and scripts are doomed to leave out too much. Explicit representation can't capture intelligence.

Dreyfus, H. L. & Dreyfus, S. E. 1988. Making a mind versus modeling the brain: AI at a crossroads. *Daedalus*.

History of AI (boo) and connectionism (qualified hooray). And Husserl/ Heidegger/Wittgenstein. Quite nice.

Hadley, R. F. 1991. The many uses of 'belief' in AI. *Minds and Machines* 1:55-74.

Various AI approaches to belief: syntactic, propositional/meaning-based, information, tractability, discoverability, and degree of confidence.

Haugeland, J. 1979. Understanding natural language. *Journal of Philosophy* 76:619-32. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

AI will need holism: interpretational, common-sense, situational, existential.

Kirsh, D. 1991. Foundations of AI: The big issues. *Artificial Intelligence* 47:3-30.

Identifying the dividing lines: pre-eminence of knowledge, embodiment, language-like kinematics, role of learning, uniformity of architecture.

Marr, D. 1977. Artificial intelligence: A personal view. *Artificial Intelligence* 9:37-48.

AI usually comes up with Type 2 (algorithmic) theories, when Type 1 (info processing) theories might be more useful -- at least if they exist.

McDermott, D. 1981. Artificial intelligence meets natural stupidity. In (J. Haugeland, ed) *Mind Design*. MIT Press.

Problems in AI methodology: wishful mnemonics, oversimplifying natural language concepts, and never implementing programs. Entertaining.

McDermott, D. 1987. A critique of pure reason. *Computational Intelligence* 3:151-60.

Criticism of logicism (i.e. reliance on deduction) in AI.

Nilsson, N. 1991. Logic and artificial intelligence. *Artificial Intelligence* 47:31-56.

Partridge, D. & Wilks, Y. (eds) 1990. *The Foundations of Artificial Intelligence: A Sourcebook*. Cambridge University Press.

Lots of papers on various aspects of AI methodology. Quite thorough.

Preston, B. 1993. Heidegger and artificial intelligence. *Philosophy and Phenomenological Research* 53:43-69.

On the non-represented background to everyday activity, and environmental interaction in cognition. Criticizes cognitivism, connectionism, looks at Agre/Chapman/Brooks, ethology, anthropology for support.

Pylyshyn, Z. W. 1979. Complexity and the study of artificial and human intelligence. In (M. Ringle, ed) *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

Ringle, M. (ed) 1979. *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

10 papers on philosophy of AI, psychology and knowledge representation.

Robinson, W. S. 1991. Rationalism, expertise, and the Dreyfuses' critique of AI research. *Southern Journal of Philosophy* 29:271-90.

Defending limited rationalism: i.e. a theory of intelligence below the conceptual level but above the neuronal level.

4.6

Computationalism in Cognitive Science [see also 2.2e]

Antony, L. 1997. Feeling fine about the mind. *Philosophy and Phenomenological Research* 57:381-87.

Bickhard, M. 1996. Troubles with computationalism. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Block, N. 1990. The computer model of mind. In (D. Osherson & E. Smith, eds) *An Invitation to Cognitive Science*, Vol. 3. MIT Press.

Overview of computationalism. Relationship to intentionality, LOT, etc.

Boden, M. 1984. What is computational psychology? *Proceedings of the Aristotelian Society* 58:17-35.

Bringsjord, S. 1994. Computation, among other things, is beneath us. *Minds and Machines* 4:469-88.

Bringsjord, S. & Zenzen, M. 1997. Cognition is not computation: The argument from irreversibility. *Synthese* 113:285-320.

Buller, D. J. 1993. Confirmation and the computational paradigm, or, why do you think they call it artificial intelligence? *Minds and Machines* 3:155-81.

Chalmers, D. J. 1994. A computational foundation for the study of cognition. Manuscript.

Argues for theses of computational sufficiency and computational explanation, resting on the fact that computation provides an abstract specification of causal organization. With replies to many anti-computationalist worries.

Clarke, J. 1972. Turing machines and the mind-body problem. *British Journal for the Philosophy of Science* 23:1-12.

Cummins, R. 1977. Programs in the explanation of behavior. *Philosophy of Science* 44:269-87.

Demopoulos, W. 1987. On some fundamental distinctions of computationalism. *Synthese* 70:79-96.

On analog/digital, representational/nonrepresentational, direct/indirect.

Dietrich, E. 1990. Computationalism. *Social Epistemology*.

What computationalism is, as opposed to computerism & cognitivism. Implies: intentionality isn't special, and we don't make decisions. With commentary.

Dietrich, E. 1989. Semantics and the computational paradigm in computational psychology. *Synthese* 79:119-41.

Argues that computational explanation requires the attribution of semantic content. Addresses Stich's

arguments against content, and argues that computers are not formal symbol manipulators.

Double, R. 1987. The computational model of the mind and philosophical functionalism. *Behaviorism* 15:131-39.

Dretske, F. 1985. Machines and the mental. *Proceedings and Addresses of the American Philosophical Association* 59:23-33.

Machines can't even add, let alone think, as the symbols they use aren't meaningful to them. They would need real information based on perceptual embodiment, and conceptual capacities, for meaning to play a real role.

Fetzer, J. H. 1994. Mental algorithms: Are minds computational systems? *Pragmatics and Cognition* 21:1-29.

Fodor, J. A. 1978. Computation and reduction. *Minnesota Studies in the Philosophy of Science* 9. Reprinted in *RePresentations* (MIT Press, 1980).

Fodor, J. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. MIT Press.

Garson, J. W. 1993. Mice in mirrored mazes and the mind. *Philosophical Psychology* 6:123-34.

Computationalism is false, as it can't distinguish the ability to solve a maze for the ability to solve its mirror image. An embodied computational taxonomy is needed, rather than software alone.

Harnad, S. 1996. Computation is just interpretable symbol manipulation; Cognition isn't. *Minds and Machines* 4:379-90.

Horst, S. 1996. *Symbols, Computation, and Intentionality: A Critique of the Computational Theory of Mind*. University of California Press.

Horst, S. 1999. Symbols and computation: A critique of the computational theory of mind. *Minds and Machines* 9:347-381

Mellor, D. H. 1984. What is computational psychology? II. *Proceedings of the Aristotelian Society* 58:37-53.

Mellor, D. H. 1989. How much of the mind is a computer. In (P. Slezak, ed) *Computers, Brains and Minds*. Kluwer.

Only belief is computational: rest of mind is not.

Nelson, R. 1987. Machine models for cognitive science. *Philosophy of Science*

Argues contra Pylyshyn 1984 that finite state automata are good models for cognitive science: they are semantically interpretable and process symbols.

Pollock, J. 1989. *How to Build a Person: A Prolegomenon*. MIT Press.

Pylyshyn, Z. W. 1980. Computation and cognition: Issues in the foundation of cognitive science. *Behavioral and Brain Sciences* 3:111-32.

Pylyshyn, Z. W. 1984. *Computation and Cognition*. MIT Press.

A thorough account of the symbolic/computational view of cognition.

Pylyshyn, Z. W. 1978. Computational models and empirical constraints. *Behavioral and Brain Sciences* 1:98-128.

Pylyshyn, Z. W. 1989. Computing and cognitive science. In (M. Posner, ed) *Foundations of Cognitive Science*. MIT Press.

An overview of the computational view of mind. On symbols, levels, control structures, levels of correspondence for computational models, and empirical methods for determining degrees of equivalence.

Shapiro, S. C. 1995. Computationalism. *Minds and Machines* 5:467-87.

Sterelny, K. 1989. Computational functional psychology: problems and prospects. In (P. Slezak, ed) *Computers, Brains and Minds*. Kluwer.

Various points on pros and cons of computational psychology.

Tibbetts, P. 1996. Residual dualism in computational theories of mind. *Dialectica* 50:37-52.

4.7 Computation and Physical Systems

Boyle, C. F. 1994. Computation as an intrinsic property. *Minds and Machines* 4:451-67.

Chalmers, D. J. 1994. On implementing a computation. *Minds and Machines* 4:391-402.

Gives an account of what it is for a physical system to implement a computation: the causal structure of the system must mirror the formal structure of the computation. Answers objections by Searle and others.

Chalmers, D. J. 1996. Does a rock implement every finite-state automaton? *Synthese* 108:309-33.

Argues that Putnam's "proof" that every ordinary open system implements every finite automaton is fallacious. It can be patched up, but an appropriate account of implementation resists these difficulties.

Chrisley, R. L. 1994. Why everything doesn't realize every computation. *Minds and Machines* 4:403-20.

Cleland, C. 1993. Is the Church-Turing thesis true? *Minds and Machines* 3:283-312.

Many physically realized functions can't be computed by Turing machines: e.g. "mundane procedures" and continuous functions. So the C-T thesis is false of these, and maybe even of number-theoretic functions.

Cleland, C. E. 1995. Effective procedures and computable functions. *Minds and Machines* 5:9-23.

Copeland, B. J. 1996. What is computation? *Synthese* 108:335-59.

Endicott, R. P. 1996. Searle, syntax, and observer-relativity. *Canadian Journal of Philosophy* 26:101-22.

Goel, V. 1991. Notationality and the information processing mind. *Minds and Machines* 1:129-166.

Adapts Goodman's notational systems to explicate computational information processing. What is/isn't a physical notational system (e.g. LOT, symbol systems, connectionism) and why. How to reconcile notational/mental content?

Hardcastle, V. G. 1995. Computationalism. *Synthese* 105:303-17.

Pragmatic factors are vital in connecting the theory of computation with empirical theory, and particularly in determining whether a given system counts as performing a given computation.

Horsten, L. 1995. The Church-Turing thesis and effective mundane procedures. *Minds and Machines* 5:1-8.

MacLennan, B. 1994. "Words lie in our way". *Minds and Machines* 4:421-37.

Miscevic, N. 1996. Computationalism and the Kripke-Wittgenstein paradox. *Proceedings of the Aristotelian Society* 96:215-29.

Scheutz, M. 1999. When physical systems realize functions. *Minds and Machines* 9:161-196.

Searle, J. R. 1990. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association* 64:21-37.

Syntax isn't intrinsic to physics, so computational ascriptions are assigned by observer. Syntax has no

causal powers. Brain doesn't process information.

Shagrir, O. 1997. Two dogmas of computationalism. *Minds and Machines* 7:321-44.

Stabler, E. 1987. Kripke on functionalism and automata. *Synthese* 70:1-22.

Disputes Kripke's argument that there is no objective way of determining when a system computes a given function, due to infinite domains and unreliability. Stipulating normal background conditions is sufficient.

Suber, P. 1988. What is software? *Journal of Speculative Philosophy* 2:89-119.

4.8

Philosophy of AI, Misc

Bergadano, F. 1993. Machine learning and the foundations of inductive inference. *Minds and Machines* 3:31-51.

Button, G. , Coulter, J. , Lee, J. R. E. & Sharrock, W. 1995. *Computers, Minds, and Conduct*. Polity Press.

Fetzer, J. H. 1990. *Artificial Intelligence: Its Scope and Limits*. Kluwer.

Gips, J. 1994. Toward the ethical robot. In (K. M. Ford, C. Glymour, & P. Hayes, eds) *Android Epistemology*. MIT Press.

Haugeland, J. (ed) 1981. *Mind Design*. MIT Press.

12 papers on the foundations of AI and cognitive science.

Hayes, P. J. , Ford, K. M. , & Adams-Webber, J. R. 1994. Human reasoning about artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence* 4:247-63. Reprinted in (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

Krellenstein, M. 1987. A reply to 'Parallel computation and the mind-body problem'. *Cognitive Science* 11:155-7.

Thagard 1986 is wrong: speed and the like make no fundamental difference. With Thagard's reply: it makes a difference in practice, if not in principle.

Moody, T. C. 1993. *Philosophy and Artificial Intelligence*. Prentice-Hall.

Preston, B. 1991. AI, anthropocentrism, and the evolution of "intelligence.". *Minds and Machines* 1:259-277.

Robinson, W. S. 1992. *Computers, Minds, and Robots*. Temple University Press.

Russell, S. 1991. Inductive learning by machines. *Philosophical Studies* 64:37-64.

A nice paper on the relationship between techniques of theory formation from machine learning and philosophical problems of induction and knowledge.

Rychlak, J. F. 1991. *Artificial Intelligence and Human Reason: A Teleological Critique*. Columbia University Press.

Sloman, A. 1978. *The Computer Revolution in Philosophy*. Harvester.

All about how the computer should change the way we think about the mind.

Thagard, P. 1986. Parallel computation and the mind-body problem. *Cognitive Science* 10:301-18.

Parallelism does make a difference. Some somewhat anti-functionalist points.

Thagard, P. 1990. Philosophy and machine learning. *Canadian Journal of Philosophy* 20:261-76.

Thagard, P. 1991. Philosophical and computational models of explanation. *Philosophical Studies* 64:87-104.

A comparison of philosophical and AI approaches to explanation: deductive, statistical, schematic, analogical, causal, and linguistic.

Winograd, T. & Flores, F. 1987. *Understanding Computers and Cognition*. Addison-Wesley.

Part 5: Philosophy of Psychology [561]

Part of [Contemporary Philosophy of Mind: An Annotated Bibliography](#).

Compiled by [David J. Chalmers](#), Department of Philosophy, University of Arizona, Tucson AZ 85721. E-mail: chalmers@arizona.edu.

- 5.1 [Issues in Cognitive Science](#) [230]
 - 5.1a [Nativism](#) (Chomsky, etc) [37]
 - 5.1b [Modularity](#) (Fodor, etc) [30]
 - 5.1c [Mental Imagery](#) [45]
 - 5.1d [Rationality](#) [31]
 - 5.1e [Embodiment](#) [31]
 - 5.1f [Animal Cognition](#) [56]
- 5.2 [Aspects of Mind](#) [224]
 - 5.2a [Pain and Pleasure](#) [40]
 - 5.2b [Emotions](#) [46]
 - 5.2c [Dreams](#) [21]
 - 5.2d [Memory](#) [32]
 - 5.2e [Color](#) [85]
- 5.3 [Philosophy of Psychology and Cognitive Science](#) [80]
 - 5.3a [Psychological Laws](#) [13]
 - 5.3b [Psychology and Neuroscience](#) [33]
 - 5.3c [Explanation in Cognitive Science](#) [17]
 - 5.3d [Philosophy of Cognitive Science, Misc](#) [17]
- 5.4 [Philosophy of Mind](#) (General) [27]

5.1 Issues in Cognitive Science

5.1a

Nativism (Chomsky, etc)

Ariew, A. 1996. Innateness and canalization. *Philosophy of Science Supplement* 63:19-27.

Atherton, M. & Schwarz, R. 1974. Linguistic innateness and its evidence. *Journal of Philosophy* 71:6.

Chomsky, N. 1967. Recent contributions to the theory of innate ideas. *Synthese* 17:2-11.

Chomsky, N. 1969. Linguistics and philosophy. In (S. Hook, ed) *Language and Philosophy*. New York University Press.

Reply to Putnam 1967: Putnam underestimates complexity of grammar, etc.

Chomsky, N. 1975. On cognitive capacity. In *Reflections on Language*. Pantheon Books.

Chomsky, N. 1980. Discussion of Putnam's comments. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

Chomsky, N. & Fodor, J. A. 1980. The inductivist fallacy. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

Churchland, P. S. 1978. Fodor on language learning. *Synthese* 38:149-59.

Cowie, F. 1997. The logical problem of language acquisition. *Synthese* 111:17-51.

Cowie, F. 1998. *What's Within?* Oxford University Press.

Cowie, F. 1998. Mad dog nativism. *British Journal for the Philosophy of Science* 49:227-252.

Cummins, D. D. 1996. Evidence for the innateness of deontic reasoning. *Mind and Language* 11:160-90.

De Rosa, R. 2000. On Fodor's claim that classical empiricists and rationalists agree on the innateness of ideas. *Protosociology* 14:240-269.

Harman, G. 1969. Linguistic competence and empiricism. In (S. Hook, ed) *Language and Philosophy*. New York University Press.

Fodor, J. A. , Bever, T. & Garrett, M. 1974. The specificity of language skills. In *The Psychology of Language*. McGraw-Hill.

Fodor, J. A. 1980. Reply to Putnam. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

Fodor, J. A. 1981. The present status of the innateness controversy. In *Representations*. MIT Press.

Concepts are undefinable, so primitive, so innate (plus gloss).

Fodor, J. A. 1980. On the impossibility of acquiring 'more powerful' structures. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard

University Press.

Katz, J. 1966. Innate ideas. In *The Philosophy of Language*. Harper & Row.

Overview; poverty of stimulus, unobservable features => rationalism.

Kaye, L. J. 1993. Are most of our concepts innate? *Synthese* 2:187-217.

Mehler, J. & Fox, R. (eds) 1985. *Neonate Cognition: Beyond the Blooming Buzzing Confusion*. Lawrence Erlbaum.

Piattelli-Palmarini, M. (ed) 1980. *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*. Harvard University Press.

An excellent collection of papers & responses by Piaget, Chomsky and others.

Piattelli-Palmarini, M. 1986. The rise of selective theories: A case study and some lessons from immunology. In (W. Demopoulos, ed) *Language Learning and Concept Acquisition*. Ablex.

Piattelli-Palmarini, M. 1989. Evolution, selection, and cognition: From learning to parameter setting in biology and in the study of language. *Cognition* 31:1-44.

Why learning is selective and not instructive. Biological analogies, linguistic evidence. Dispense with "learning" as a scientific term.

Pitt, D. 2000. Nativism and the theory of content. *Protosociology* 14:222-239.

Putnam, H. 1967. The 'Innateness Hypothesis' and explanatory models in linguistics. *Synthese* 17:12-22. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Contra nativism: disputes (1) surprising universals (2) explanation of universals (3) ease of learning (4) relevance of IQ-independence.

Putnam, H. 1980. What is innate and why. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

Putnam, H. 1980. Comments on Chomsky's and Fodor's replies. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

Ramsey, W. & Stich, S. P. 1990. Connectionism and three levels of nativism. *Synthese* 82:177-205.

Identifies minimal nativism vs anti-empiricism vs rationalism. Considers the relevance of connectionist

networks. Some nativist arguments may survive.

Samet, J. 1986. Troubles with Fodor's nativism. *Midwest Studies in Philosophy* 10:575-594.

Concepts can be acquired without being learned by symbol-manipulation.

Samet, J. & Flanagan, O. J. 1989. Innate representations. In (S. Silvers, ed) *Rerepresentation*. Kluwer.

Sampson, G. 1978. Linguistic universals as evidence for empiricism. *Journal of Linguistics*.

Explain universals via Popper/Simon empirical model.

Samuels, R. 1998. What brains won't tell us about the mind: A critique of the neurobiological argument against representational nativism. *Mind and Language* 13:548-570.

Schwartz, R. 1995. Is mathematical competence innate? *Philosophy of Science* 62:227-40.

Sterelny, K. 1989. Fodor's nativism. *Philosophical Studies* 55:119-41.

Stich, S. P. (ed) 1975. *Innate Ideas*. University of California Press.

Stich, S. P. 1979. Between Chomskian rationalism and Popperian empiricism. *British Journal for the Philosophy of Science* 30:329-47.

Can take middle ground. Anti-empiricism doesn't imply rationalism.

5.1b Modularity (Fodor, etc)

Appelbaum, I. 1998. Fodor, modularity, and speech perception. *Philosophical Psychology* 11:317-330.

Arbib, M. 1989. Modularity, schemas and neurons: A critique of Fodor. In (P. Slezak, ed) *Computers, Brains and Minds*. Kluwer.

Against Fodor: modules are smaller, interact strongly, not domain-specific.

Bennett, L. J. 1990. Modularity of mind revisited. *British Journal for the Philosophy of Science* 41:429-36.

Remarks on Shanon and Fodor.

Browne, D. 1996. Cognitive versatility. *Minds and Machines* 6:507-23.

Bruner, J. 1957. On perceptual readiness. *Psychological Review* 65:14-21.

Overview of the original studies on top-down effects in perception.

Cam, P. 1988. Modularity, rationality, and higher cognition. *Philosophical Studies* 53:279-94.

Cam, P. 1990. Insularity and the persistence of perceptual illusion. *Analysis* 50:231-5.

Chien, A. J. 1996. Why the mind may not be modular. *Minds and Machines* 6:1-32.

Churchland, P. M. 1979. *Scientific Realism and the Plasticity of Mind*. Cambridge University Press.

Our perception is deeply theory-laden, and potentially very plastic.

Churchland, P. M. 1988. Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor. *Philosophy of Science* 55:167-87. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Contra Fodor 1984: observation is theory-laden (built-in or not); supported by neurophysiological evidence; perceptual systems have long-term plasticity.

Currie, G. & Sterelny, K. 2000. How to think about the modularity of mind-reading. *Philosophical Quarterly* 50:145-160.

DesAutels, P. 1995. Two types of theories: The impact of Churchland's perceptual plasticity. *Philosophical Psychology* 8:25-33.

Fodor, J. A. 1983. *The Modularity of Mind*. MIT Press.

Perception happens in informationally encapsulated, domain-specific modules. Central systems aren't encapsulated, and so may be impossible to understand.

Fodor, J. A. 1985. *Precis of The modularity of mind*. *Behavioral and Brain Sciences* 8:1-42. Reprinted in *A Theory of Content and Other Essays* (MIT Press, 1990).

Summary of MOM (with commentary and reply in the BBS printing).

Fodor, J. A. 1986. The modularity of mind. In (Z. Pylyshyn, ed) *Meaning and Cognitive Structure*. Ablex.

Informal discussion of modularity. With commentaries by Fahlman, Caplan.

Fodor, J. A. 1984. Observation reconsidered. *Philosophy of Science* 51:23-43. Reprinted in *A Theory of*

Content and Other Essays (MIT Press, 1990).

Argues for an observation/theory distinction, and against belief affecting perception.

Fodor, J. A. 1988. A reply to Churchland's 'Perceptual plasticity and theoretical neutrality'. *Philosophy of Science* 55:188-98. Reprinted in *A Theory of Content and Other Essays* (MIT Press, 1990).

Churchland is up the creek without a paddle.

Fodor, J. A. 1989. Why should the mind be modular? In (A. George, ed) *Reflections on Chomsky*. Blackwell. Reprinted in *A Theory of Content and Other Essays* (MIT Press, 1990).

Garfield, J. (ed) 1987. *Modularity in Knowledge Representation and Natural-Language Understanding*. MIT Press.

A collection of papers on modularity in language and vision.

Gray, R. 2001. Cognitive modules, synaesthesia and the constitution of psychological natural kinds. *Philosophical Psychology* 14:65-82.

Meyering, T. C. 1994. Fodor's modularity: A new name for an old dilemma. *Philosophical Psychology* 7:39-62.

Olsson, E. 1997. Coherence and the modularity of mind. *Australasian Journal of Philosophy* 75:404-11.

Pylyshyn, Z. 1999. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences* 22:341-365.

Rollins, M. 1994. Deep plasticity: The encoding approach to perceptual change. *Philosophy of Science* 61:39-54.

Ross, J. 1990. Against postulating central systems in the mind. *Philosophy of Science* 57:297-312.

Fodor's arguments for unencapsulated central systems are no good; AI is possible after all.

Shanon, B. 1988. Remarks on the modularity of mind. *British Journal for the Philosophy of Science* 39:331-52.

Criticism of Fodor. Modularity is dynamic, and can be central.

Fodor, J. A. & Pylyshyn, Z. W. 1981. How direct is visual perception?: Some reflections on Gibson's 'ecological approach'. *Cognition* 9:139-96.

'Direct perception' can't correspond to anything. Perception is inferential.

Turvey, M. T. , Shaw, R. E. , Reed, E. S. , & Mace, W. M. 1981. Ecological laws of perceiving and acting: In Reply to Fodor and Pylyshyn. *Cognition* 9:237-304.

Ullman, S. 1980. Against direct perception. *Behavioral and Brain Sciences* 3:333-81.

Vaina, L. M. 1990. What and where in the human visual system: Two hierarchies of visual modules. *Synthese* 83:49-91.

5.1c Mental Imagery [see also 6.2j]

Anderson, J. R. 1978. Arguments concerning representations for mental imagery. *Psychological Review* 85.

Audi, R. 1978. The ontological status of mental images. *Inquiry* 21:348-61.

Blachowicz, J. 1997. Analog representation beyond mental imagery. *Journal of Philosophy* 94:55-84.

Block, N. (ed) 1981. *Imagery*. MIT Press.

Block, N. 1983. Mental pictures and cognitive science. *Philosophical Review* 93:499-542. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Block, N. 1983. The photographic fallacy in the debate about mental imagery. *Nous* 17:651-62.

Brown, R. & Herrnstein, R. 1981. Icons and images. In (N. Block, ed) *Imagery*. MIT Press.

Cam, P. 1987. Propositions about images. *Philosophy and Phenomenological Research* 48:335-8.

Candlish, S. 1975. Mental images and pictorial properties. *Mind* 84:260-2.

Chambers, D. & Reisberg, D. 1985. Can mental images be ambiguous?' *Journal of Experimental Psychology: Human Perception and Performance* 11:317-28.

Chambers, D. & Reisberg, D. 1992. What an image depicts depends on what an image means. *Cognitive Psychology* 24:145-74.

Dennett, D. C. 1978. Two approaches to mental images. In *Brainstorms*. MIT Press.

- Dennett, D. C. 1968. The nature of images and the introspective trap. In *Content and Consciousness*. Routledge and Kegan Paul. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).
- Farah, M. J. 1988. Is visual imagery really visual: Some overlooked evidence from neuropsychology. *Psychological Review* 95:307-17.
- Finke, R. A. 1989. *Principles of Mental Imagery*. MIT Press.
- Fodor, J. A. 1975. Imagistic representation. In *The Language of Thought*. Harvard University Press.
- Glasgow, J. I. 1993. The imagery debate revisited: A computational perspective. *Computational Intelligence* 9:310-33.
- Hannay, A. 1971. *Mental Images: A Defense*. Allen & Unwin.
- Hannay, A. 1973. To see a mental image. *Mind* 82:161-262.
- Kind, A. 2001. Putting the image back in imagination. *Philosophy and Phenomenological Research* 62:85-110.
- Kosslyn, S. M. & Pomerantz, J. 1977. Imagery, propositions and the form of internal representations. *Cognitive Psychology* 9:52-76. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).
- Kosslyn, S. M. 1981. The medium and the message in mental imagery: A theory. In (N. Block, ed) *Imagery*. MIT Press.
- Kosslyn, S. M. , Pinker, S. , Schwartz, S. & Smith, G. 1979. On the demystification of mental imagery. *Behavioral and Brain Sciences* 2:535-81.
- Kosslyn, S. M. 1980. *Image and Mind*. Harvard University Press.
- Kosslyn, S. M. 1994. *Image and Brain: The Resolution of the Imagery Debate*. MIT Press.
- Maloney, J. C. 1984. Mental images and cognitive theory. *American Philosophical Quarterly* 21:237-47.
- Morris, P. E. & Hampson, P. J. 1983. *Imagery and Consciousness*. Academic Press.
- Mortensen, C. 1989. Mental images: Should cognitive science learn from neurophysiology? In (P. Slezak, ed) *Computers, Brains and Minds*. Kluwer.

- Pylyshyn, Z. W. 1973. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psych Bull* 80:1-24.
- Pylyshyn, Z. W. 1978. Imagery and artificial intelligence. In (W. Savage, ed) *Perception and Cognition*. University of Minnesota Press. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).
- Pylyshyn, Z. W. 1981. The imagery debate: Analog media vs. tacit knowledge. *Psychological Review* 88:16-45. Reprinted in Block 1981.
- Reisberg, D. & Chambers, D. 1991. Neither pictures nor propositions: What can we learn from a mental image? *Canadian Journal of Psychology* 45:336-52.
- Rey, G. 1981. What are mental images? In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol. 2. Harvard University Press.
- Richardson, A. 1969. *Mental Imagery*. Routledge.
- Rollins, M. 1989. *Mental Imagery: On the Limits of Cognitive Science*. Yale University Press.
- Russow, L. 1985. Dennett, mental images and images in context. *Philosophy and Phenomenological Research* 45:581-94.
- Schwartz, R. 1980. Imagery: There is more to it than meets the eye. *Philosophy of Science Association* 1980.
- Shepard, R. & Cooper, L. 1982. *Mental Images and their Transformations*. MIT Press.
- Shier, D. 1997. How can pictures be propositions? *Ratio* 10:65-75.
- Sterelny, K. 1986. The imagery debate. *Philosophy of Science* 53:560-83. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).
- Thomas, N. J. T. 1997. Are theories of imagination theories of imagery? Manuscript.
- Tye, M. 1984. The debate about mental imagery. *Journal of Philosophy* 81:678-91.
- Tye, M. 1988. The picture theory of images. *Philosophical Review*.
- Tye, M. 1991. *The Imagery Debate*. MIT Press,

Wright, E. 1983. Inspecting images. *Philosophy* 58:57-72.

5.1d Rationality

Biro, J. & Ludwig, K. 1994. Are there more than minimal a priori limits on irrationality? *Australasian Journal of Philosophy* 72:89-102.

Cherniak, C. 1986. *Minimal Rationality*. MIT Press.

Cherniak, C. 1981. Minimal rationality. *Mind* 90:161-83.

Cherniak, C. 1983. Rationality and the structure of memory. *Synthese* 57:163-86.

Cohen, L. J. 1979. On the psychology of prediction: Whose is the fallacy? *Cognition* 7:385-407.

Cohen, L. J. 1980. Whose is the fallacy? A rejoinder to Daniel Kahneman and Amos Tversky. *Cognition* 8:89-92.

Cohen, L. J. 1981. Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*.

Cohen, L. J. 1986. *The Dialogue of Reason*. Cambridge University Press.

Cook, K. S. & Levi, M. 1990. *The Limits of Rationality*. University of Chicago Press.

Davidson, D. 1985. Incoherence and irrationality. *Dialectica* 39:345-54.

Davidson, D. 1995. Could there be a science of rationality? *International Journal of Philosophical Studies* 3:1-16.

Feldman, R. 1988. Rationality, reliability, and natural selection. *Philosophy of Science* 55:218-27.

Fetzer, J. H. 1990. Evolution, rationality and testability. *Synthese* 82:423-39.

Gardner, S. 1996. *Irrationality and the Philosophy of Psychoanalysis*. Cambridge University Press.

Harman, G. 1986. *Change in View*. MIT Press.

Holt, L. 1999. Rationality is still hard work: Some further notes on the disruptive effects of deliberation. *Philosophical Psychology* 12:215-219.

- Kahneman, D. , Slovic, P. & Tversky, A. (eds) 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.
- Kahneman, D. & Tversky, A. 1979. On the interpretation of intuitive probability: A reply to Jonathan Cohen. *Cognition* 7:409-11.
- Manktelow, K. & Over, D. 1987. Reasoning and rationality. *Mind and Language* 2:199-219.
- Mele, A. R. 1987. *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford University Press.
- Nisbett, R. & Ross, L. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall.
- Nozick, R. 1993. *The Nature of Rationality*. Princeton University Press.
- Reiner, R. 1995. Arguments against the possibility of perfect rationality. *Minds and Machines* 5:373-89.
- Rust, J. 1990. Delusions, irrationality, and cognitive science. *Philosophical Psychology*.
- Scholl, B. J. 1997. Reasoning, rationality, and architectural resolution. *Philosophical Psychology* 10:451-470.
- Scott-Kakures, D. 1996. Self-deception and internal irrationality. *Philosophy and Phenomenological Research* 56:31-56.
- Sober, E. 1981. The evolution of rationality. *Synthese* 46:95-120.
- Stein, E. 1994. Rationality and reflective equilibrium. *Synthese* 99:137-72.
- Stein, E. 1996. *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford University Press.
- Stich, S. P. 1985. Could man be an irrational animal? *Synthese* 64:115-35.
- Wason, P. 1966. Reasoning. In (Foss, ed) *New Horizons in Psychology*. Penguin.

5.1e

Embodiment [see also 2.2]

- Agre, P. 1995. Computation and embodied agency. *Informatica* 19:527-35.
- Ballard, D. 1991. Animate vision. *Artificial Intelligence* 48:57-86.
- Beer, R. 1995. A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* 72:173-215.
- Bermudez, J. L. , Marcel, A. , & Eilan, N. (eds) 1995. *The Body and the Self*. MIT Press.
- Buckley, J. & Hall, L. 1999. Self-knowledge and embodiment. *Southwest Philosophy Review* 15.
- Chrisley, R. L. 1994. Taking embodiment seriously: Nonconceptual content and robotics. In (K. M. Ford, C. Glymour, & P. Hayes, eds) *Android Epistemology*. MIT Press.
- Clark, A. 1987. Being there: Why implementation matters to cognitive science. *AI Review* 1:231-44.
- On the importance of embodiment of systems in cognition.
- Clark, A. 1995. Moving minds: Situating content in the service of real-time success. *Philosophical Perspectives* 9:89-104.
- Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again*. MIT Press.
- Clark, A. & Chalmers, D. J. 1998. The extended mind. *Analysis* 58:7-19.
- Advocates a sort of "active externalism", based on the role of the environment in actively driving cognition. Beliefs can extend into an agent's immediate environment (e.g. a notebook) in this way.
- Clark, A. 2001. Reasons, robots and the extended mind. *Mind and Language* 16:121045.
- Cussins, A. 1992. Content, embodiment, and objectivity: The theory of cognitive trails. *Mind* 101:651-88.
- Gibson, J. J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.
- Godfrey-Smith, P. 1996. *Complexity and the Function of Mind in Nature*. Cambridge University Press.
- Haugeland, J. 1993. Mind embodied and embedded. In (Y. Houng & J. Ho, eds) *Mind and Cognition:1993 International Symposium*. Academia Sinica.

Argues that the mind is not just embedded but intimately intermingled with the world. With some systems-

theoretic arguments arguing against a determinate interface. Mind is not an inner realm.

Hendriks-Jansen, H. 1996. *Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought*. MIT Press.

Hutchins, E. 1995. *Cognition in the Wild*. MIT Press.

Johnson, M. L. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press.

Johnson, M. L. 1995. Incarnate mind. *Minds and Machines* 5:533-45.

Kirsh, D. & Maglio, P. 1995. On distinguish epistemic from pragmatic action. *Cognitive Science* 18:513-49.

Loren, L. A. & Dietrich, E. 1997. Merleau-Ponty, embodied cognition, and the problem of intentionality. *Cybernetics and Systems* 28:345-58.

Losonsky, M. 1995. Emdedded systems vs. individualism. *Minds and Machines* 5:357-71.

McClamrock, R. 1995. *Existential Cognition: Computational Minds in the World*. University of Chicago Press.

O'Regan, K. 1992. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology* 46:461-88.

Rosenschein, S. J. & Kaelbling, L. P. 1995. A situated view of representation and control. *Artificial Intelligence* 73:149-73.

van Gelder, T. 1993. The distinction between mind and cognition. In (Y. Houng & J. Ho, eds) *Mind and Cognition: 1993 International Symposium*. Academia Sinica.

Argues against the contemporary "Cartesian" view of mind as an ontologically homogeneous inner representational realm that causes behavior, arguing for a holistic embodied view instead. Mind is therefore safe from elimination.

Varela, F. , Thompson, E. & Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Vera, A. H. & Simon, H. A. 1993. Situated action: A symbolic interpretation. *Cognitive Science* 17:7-48.

Wells, A. 1996. Situated action, symbol systems and universal computation. *Minds & Machines* 6:33-46.

Wilkerson, W. S. 1999. From bodily motions to bodily intentions: the perception of bodily activity. *Philosophical Psychology* 12:61-77.

Zhang, J. & Norman, D. 1994. Representations in distributed cognitive tasks. *Cognitive Science* 18:87-122.

5.1f Animal Cognition [see also 6.4c, 6.4d]

Allen, C. 1997. Animal cognition and animal minds. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Allen, C. 1999. Animal concepts revisited: the use of self-monitoring as an empirical approach. *Erkenntnis* 51:537-544.

Allen, C. & Bekoff, M. 1992. On aims and methods of cognitive ethology. *Philosophy of Science Association* 1992, 2:110-24.

Allen, C. & Bekoff, M. 1995. Cognitive ethology and the intentionality of animal behavior. *Mind and Language* 10:313-328.

Allen, C. & Bekoff, M. 1997. *Species of Mind: The Philosophy and Biology of Cognitive Ethology*. MIT Press.

Allen, C. & Hauser, M. 1991. Concept attribution in nonhuman animals: Theoretical and methodological problems in ascribing complex mental processes. *Philosophy of Science* 58:221-40. Reprinted in Allen & Jamison 1996.

Bateson, P. P. G. & Klopfer, P. H. 1991. *Perspectives in Ethology, Volume 9: Human Understanding and Animal Awareness*. Plenum Press.

Beer, C. G. 1992. Conceptual issues in cognitive ethology. *Advances in the Study of Behavior* 21:69-109.

Bekoff, M. & Jamieson, D. (eds) 1996. *Readings in Animal Cognition*. MIT Press.

Bekoff, M. 1999. Social cognition: Exchanging and sharing information on the run. *Erkenntnis* 51:617-632.

Cheney, D. L. & Seyfarth, R. M. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. University of Chicago Press.

- Clark, S. R. L. 1987. The description and evaluation of animal emotion. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.
- Cockburn, D. 1994. Human beings and giant squids (on ascribing human sensations and emotions to non-human creatures). *Philosophy* 69:135-50.
- Crisp, R. 1996. Evolution and psychological unity. In (M. Bekoff & D. Jamieson, eds) *Readings in Animal Cognition*. MIT Press.
- Davidson, D. 1982. Rational animals. *Dialectica* 36:317-28.
- Dawkins, M. S. 1987. Minding and mattering. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.
- Dawkins, M. S. 1990. From an animal's point of view: Motivation, fitness, and animal welfare. *Behavioral and Brain Sciences*.
- Dennett, D. C. 1983. Intentional systems in cognitive ethology: The 'Panglossian paradigm' defended. *Behavioral and Brain Sciences* 6:343-90. Reprinted in *The Intentional Stance* (MIT Press, 1987).
- Dennett, D. C. 1989. Cognitive ethology: Hunting for bargains or a wild goose chase? In (Montefiore, ed) *Goals, No-Goals and Own Goals*. Unwin Hyman.
- Dennett, D. C. 1995. Do animals have beliefs? In (H. Roitblat & J. Meyer, eds) *Comparative Approaches to Cognitive Science*. MIT Press.
- Dennett, D. C. 1996. *Kinds of Minds*. Basic Books.
- Dreckmann, F. 1999. Animal beliefs and their contents. *Erkenntnis* 51:597-615.
- Dupre, J. 1996. The mental lives of nonhuman animals. In (M. Bekoff & D. Jamieson, eds) *Readings in Animal Cognition*. MIT Press.
- Fellows, R. 2000. Animal belief. *Philosophy* 75:587-599.
- Gaita, R. 1992. Animal thoughts. *Philosophical Investigations* 15:227-44.
- Gauker, C. 1990. How to learn language like a chimpanzee. *Philosophical Psychology* 4:139-46.
- Glock, H. 2000. Animals, thoughts and concepts. *Synthese* 123:35-104.

- Gould, J. L. & Gould, C. G. 1982. The insect mind: Physics or metaphysics? In (D. Griffin, ed) *Animal Mind -- Human Mind*. Springer-Verlag.
- Gould, J. L. & Gould, C. G. 1994. *The Animal Mind*. Scientific American Library.
- Griffin, D. R. (ed) 1982. *Animal Mind -- Human Mind*. Springer-Verlag.
- Griffin, D. R. 1984. *Animal Thinking*. Harvard University Press.
- Griffin, D. R. 1992. *Animal Minds*. University of Chicago Press.
- Harrison, P. 1991. Do animals feel pain? *Philosophy* 66:25-40.
- Heil, J. 1982. Speechless brutes. *Philosophy and Phenomenological Research* 42:400-406.
- Hendrichs, H. 1999. Different roots of human intentionality in mammalian mentality. *Erkenntnis* 51:649-668.
- Malcolm, N. 1973. Thoughtless brutes. *Proceedings and Addresses of the American Philosophical Association* 46:5-20.
- Nelson, J. 1983. Do animals propositionally know? Do they propositionally believe? *American Philosophical Quarterly* 20:149-60.
- Premack, D. & Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 4:515-629.
- Premack, D. 1986. *Gavagai! or the Future History of the Animal Language Controversy*. MIT Press.
- Proust, J. 1999. Mind, space and objectivity in non-human animals. *Erkenntnis* 51:545-562.
- Radner, D. 1993. Directed action and animal communication. *Ration* 6:135-54.
- Radner, D. 1999. Mind and function in animal communication. *Erkenntnis* 51:633-648.
- Ristau, C. A. (ed) 1991. *Cognitive Ethology: The Minds of Other Animals*. Lawrence Erlbaum.
- Roberts, R. C. 1996. Propositions and animal emotion. *Philosophy* 71:147-56.
- Routley, R. 1982. Alleged problems in attributing beliefs, and intentionality, to animals. *Inquiry* 24:385-417.

Savage-Rumbaugh, E. S. , Rumbaugh, D. M. , & Boysen, S. 1980. Do apes use language? *American Scientist* 68:49-61.

Savage-Rumbaugh, S. & Brakke, K. E. 1996. Animal language: Methodological and interpretative issues. In (C. Allen & D. Jamison, eds) *Readings in Animal Cognition*. MIT Press.

Sebeok, T. A. & Umiker-Sebeok, J. 1980. *Speaking of Apes: A Critical Anthology of Two-Way Communication with Man*. Plenum Press.

Smit, H. 1995. Are animal displays bodily movements or manifestations of the mind? *Behavior and Philosophy* 23:13-19.

Sorabji, R. 1992. Animal minds. *Southern Journal of Philosophy* 31:1-18.

Stephan, A. 1999. Are animals capable of concepts? *Erkenntnis* 51:583-596.

Sterelny, K. 1990. Animals and individualism. In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

Sterelny, K. 1995. Basic minds. *Philosophical Perspectives* 9:251-70.

Stich, S. P. 1978. Do animals have beliefs? *Australasian Journal of Philosophy* 57:15-28.

Wilder, H. 1996. Interpretative cognitive ethology. In (C. Allen & D. Jamison, eds) *Readings in Animal Cognition*. MIT Press.

Wilson, M. D. 1995. Animal ideas. *Proceedings and Addresses of the American Philosophical Association* 69:7-25.

5.2 Aspects of Mind

5.2a

Pain and Pleasure

Aydede, M. 2000. An analysis of pleasure vis-a-vis pain. *Philosophy and Phenomenological Research* 61:537-570.

Beardman, S. 2000. The choice between current and retrospective evaluations of pain. 13:97-110.

- Blum, A. 1991. A note on pleasure. *Journal of Value Inquiry* 25:367-70.
- Chapman, C. R. 2000. Pain and folk theory. *Brain and Mind* 1:209-222.
- Conee, E. 1984. A defense of pain. *Philosophical Studies* 46:239-48.
- Cowan, J. 1968. *Pleasure and Pain: A Study in Philosophical Psychology*. Macmillan.
- Dartnall, T. 2001. The pain problem. *Philosophical Psychology* 14:95-102.
- Dennett, D. C. 1978. Why you can't make a computer that feels pain. *Synthese* 38. Reprinted in *Brainstorms* (MIT Press, 1978).
- Douglas, G. 1998. Why pains are not mental objects. *Philosophical Studies* 91:127-148.
- Edwards, R. 1975. Do pleasures and pains differ qualitatively? *Journal of Value Inquiry* 9:270-81.
- Garfield, J. L. 2001. Pain deproblematized. *Philosophical Psychology* 14:103-7.
- Gillett, G. 1991. The neurophilosophy of pain. *Philosophy* 66:191-206.
- Goldstein, I. 1980. Why people prefer pleasure to pain. *Philosophy* 55.
- Goldstein, I. 1989. Pleasure and pain: unconditional intrinsic values. *Philosophy and Phenomenological Research*.
- Goldstein, I. 1999. Intersubjective properties by which we specify pain, pleasure, and other kinds of mental states. *Philosophy*.
- Graham, G. & Stephens, G. 1987. Minding your P's and Q's: Pain and sensible qualities. *Nous* 21:395-405.
- Grahek, N. 1991. Objective and subjective aspects of pain. *Philosophical Psychology* 4:249-66.
- Grahek, N. 1995. The sensory dimension of pain. *Philosophical Studies* 79:167-84.
- Gustafson, D. 1995. Belief in pain. *Consciousness and Cognition* 4:323-45.
- Gustafson, D. 2000. On the supposed utility of a folk theory of pain. *Brain and Mind* 1:223-228.
- Hall, R. J. 1989. Are pains necessarily unpleasant? *Philosophy and Phenomenological Research* 49:643-

59.

Hardcastle, V. G. 1997. When a pain is not. *Journal of Philosophy* 94:381-409.

Hardcastle, V. G. 2000. *The Myth of Pain*. MIT Press.

Kaufman, R. 1985. Is the concept of pain incoherent? *Southern Journal of Philosophy* 23:279-84.

Langsam, H. 1995. Why pains are mental objects. *Journal of Philosophy* 6:303-13.

Momeyer, R. 1975. Is pleasure a sensation? *Philosophy and Phenomenological Research* 36:113-21.

Morris, K. J. 1996. Pain, injury, and first/third-person asymmetry. *Philosophy and Phenomenological Research* 56:125-56.

Nelkin, N. 1986. Pains and pain sensations. *Journal of Philosophy* 83:129-48.

Nelkin, N. 1994. Reconsidering pain. *Philosophical Psychology* 7:325-43.

Newton, N. 1989. On viewing pain as a secondary quality. *Nous* 23:569-98.

Pitcher, G. 1970. The awfulness of pain. *Journal of Philosophy* 48.

Pitcher, G. 1970. Pain perception. *Philosophical Review* 74:368-93.

Puccetti, R. 1975. Is pain necessary? *Philosophy* 50:259-69.

Quinn, W. 1968. Pleasure -- disposition or episode? *Philosophy and Phenomenological Research* 28:578-86.

Rachels, S. 2000. Is unpleasantness intrinsic to unpleasant experiences? *Philosophical Studies* 99:187-210.

Rachlin, H. 1985. Pain and behavior. *Behavioral and Brain Sciences* 8:43-83.

Resnik, D. 2000. Pain as a folk psychological concept: A clinical perspective. *Brain and Mind* 1:193-207.

Sufka, K. J. & Lynch, M. P. 2000. Sensations and pain processes. *Philosophical Psychology* 13:299-311.

Tye, M. 1995. A representational theory of pains and their phenomenal character. *Philosophical Perspectives* 9:223-39.

Williams, B. 1959. Pleasure and belief. *Proceedings of the Aristotelian Society*.

5.2b

Emotions

Addis, L. 1995. The ontology of emotion. *Southern Journal of Philosophy* 33:261-78.

Arregui, J. V. 1996. On the intentionality of moods: Phenomenology and linguistic analysis. *American Catholic Philosophical Quarterly* 70:397-411.

Bedford, E. 1957. Emotions. *Proceedings of the Aristotelian Society* 57:281-304.

Ben-Ze'ev, A. 1987. The nature of emotions. *Philosophical Studies* 52:393-409.

Ben-Ze'ev, A. 1990. Describing the emotions. *Philosophical Psychology* 3:305-17.

Ben-Ze'ev, A. 1992. Emotional and moral evaluations. *Metaphilosophy* 23:214-29.

Charland, L. C. 1995. Feeling and representing: Computational theory and the modularity of affect. *Synthese* 105:273-301.

D'Arms, J. & Jacobson, D. 2000. The moralistic fallacy: On the "appropriateness" of emotions. *Philosophy and Phenomenological Research* 61:65-90.

Davis, W. 1981. A theory of happiness. *American Philosophical Quarterly* 18:111-20.

DeLancey, C. 1997. Emotion and the computational theory of mind. In (S. O'Neill, P. McKevitt, & E. MacAogain, eds) *Two Sciences of Mind*. John Benjamins.

DeLancey, C. 1998. Real emotions. *Philosophical Psychology* 11:467-487.

de Sousa, R. 1979. The rationality of emotions. *Dialogue*.

de Sousa, R. 1987. *The Rationality of Emotion*. MIT Press.

Deigh, J. 1994. Cognitivism in the theory of emotions. *Ethics* 104:824-54.

Goldie, P. 2000. *The Emotions: A Philosophical Exploration*. Oxford University Press.

- Gordon, R. M. 1974. The aboutness of emotions. *American Philosophical Quarterly* 27:11-36.
- Gordon, R. M. 1986. The passivity of emotions. *Philosophical Review* 95:339-60.
- Gordon, R. M. 1987. *The Structure of Emotions: Investigations in Cognitive Philosophy*. Cambridge University Press.
- Green, O. 1992. *The Emotions: A Philosophical Theory*. Kluwer.
- Greenspan, P. S. 1988. *Emotions and Reasons: An Enquiry into Emotional Justification*. Routledge.
- Griffiths, P. 1989. Folk, functional and neurochemical aspects of mood. *Philosophical Psychology* 2:17-32.
- Haybron, D. M. 2001. Happiness and pleasure. *Philosophy and Phenomenological Research* 62:501-528.
- Helm, B. W. 1994. *The Significance of Emotions*. *American Philosophical Quarterly* 31:319-31.
- Irani, K. S. & Myers, G. 1983. *Emotion: Philosophical Studies*. Haven.
- Letwin, O. 1987. *Ethics, Emotion, and the Unity of the Self*. Croom Helm.
- Lormand, E. 1985. Toward a theory of moods. *Philosophical Studies* 47:385-407.
- Lyons, W. 1978. Emotions and behavior. *Philosophy and Phenomenological Research*.
- Marks, J. 1982. A theory of emotion. *Philosophical Studies* 42:227-42.
- McCullagh, C. B. 1990. The rationality of emotions and of emotional behavior. *Australasian Journal of Philosophy* 68:44-58.
- Morreal, J. 1983. Humor and emotion. *American Philosophical Quarterly* 20:297-304.
- Nash, R. A. 1989. Cognitive theories of emotion. *Nous* 23:481-504.
- Neu, J. 1971. *Emotion, Thought, and Therapy*. Cambridge University Press.
- Neu, J. 2000. *A Tear is an Intellectual Thing: The Meanings of Emotion*. Oxford University Press.
- Pugmire, D. 1994. Real emotion. *Philosophy and Phenomenological Research* 54:105-22.

- Rey, G. 1980. Functionalism and the emotions. In (A. Rorty, ed), *Explaining Emotions*. University of California Press.
- Roberts, R. C. 1995. Feeling one's emotions and knowing oneself. *Philosophical Studies* 77:319-38.
- Rorty, A. O. 1978. Explaining emotions. *Journal of Philosophy*.
- Rorty, A. O. (ed) 1980. *Explaining Emotions*. University of California Press.
- Rosenthal, D. M. 1983. Emotions and the self. In (K. Irani & G. Myers) *Emotion: Philosophical Studies*. Haven.
- Sizer, L. 2000. Towards a computational theory of mood. *British Journal for the Philosophy of Science* 51:743-770.
- Solomon, R. C. 1973. Emotion and choice. *Review of Metaphysics* 17:20-41.
- Taylor, G. 1975. Justifying the emotions. *Mind*.
- Thalberg, I. 1964. Emotion and thought. *American Philosophical Quarterly*.
- Wilkinson, S. 2000. Is 'normal grief' a mental disorder? *Philosophical Quarterly* 50:289-305.
- Wilson, J. R. S. 1972. *Emotion and Object*. Cambridge University Press.
- Wollheim, R. 1999. *On the Emotions*. Yale University Press.

5.2c

Dreams [see also 6.2I]

- Ayer, A. 1960. Professor Malcolm on dreams. *Journal of Philosophy*.
- Chappell, V. C. 1963. The concept of dreaming. *Philosophical Quarterly* 13:193-213.
- Chihara, C. 1965. What dreams are made of. *Theoria* 31:145-58.
- Curley, E. M. 1975. Dreaming and conceptual revision. *Australasian Journal of Philosophy* 53:119-41.
- Dennett, D. C. 1976. Are dreams experiences? *Philosophical Review* 73:151-71. Reprinted in

Brainstorms (MIT Press, 1978).

Dunlop, C. E. M. (ed) 1977. *Philosophical Essays on Dreaming*. Cornell University Press.

Emmett, K. 1978. Oneiric experiences. *Philosophical Studies* 34:445-50.

Flanagan, O. 1995. Deconstructing dreams: The spandrels of sleep. *Journal of Philosophy* 92:5-27.

Flanagan, O. 1996. Self-expression in sleep: Neuroscience and dreams. In *Self-Expressions*. Oxford University Press.

Flanagan, O. 2000. *Dreaming Souls: Sleep, Dreams, and the Evolution of the Conscious Mind*. Oxford University Press.

Hunter, J. 1971. Some questions about dreaming. *Mind* 80:70-92.

Hunter, J. 1983. The difference between dreaming and being awake. *Mind* 92:80-93.

Landesman, C. 1964. Dreams: Two types of explanation. *Philosophical Studies* 15:17-23.

Malcolm, N. 1962. *Dreaming*. Routledge and Kegan Paul.

Mannison, D. S. 1975. Dreaming an impossible dream. *Canadian Journal of Philosophy* 4:663-75.

Matthews, G. B. 1981. On being immoral in a dream. *Philosophy* 56:47-64.

Putnam, H. 1962. Dreaming and 'depth grammar'. In (R. Butler, ed) *Analytical Philosophy: First Series*. Oxford University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Schroeder, S. 1997. The concept of dreaming: On three theses by Malcolm. *Philosophical Investigations* 20:15-38.

Seligman, M. & Yellen, A. 1987. What is a dream? *Behavior Research and Therapy* 25:1-24.

Shanon, B. 1983. Descartes' puzzle -- An organismic approach. *Cognition and Brain Theory* 6:185-95.

Siegler, F. A. 1967. Remembering dreams. *Philosophical Quarterly* 17:14-24.

5.2d Memory [see also 2.2f, 3.7]

Arcaya, J. M. 1989. Memory and temporality: A phenomenological alternative. *Philosophical Psychology*

2:101-110.

Baier, A. 1976. Mixing memory and desire. *American Philosophical Quarterly* 13:213-20.

Ben-Zeev, A. 1986. Two approaches to memory. *Philosophical Investigations* 9:288-301.

Bergson, H. 1991. *Matter and memory*. MIT Press.

Campbell, J. 1997. The realism of memory. In (R. Heck, ed) *Language, Thought, and Logic: Essays in Honour of Michael Dummett*. Oxford University Press.

Campbell, J. 1997. The structure of time in autobiographical memory. *European Journal of Philosophy* 5:105-17.

Cherniak, C. 1983. Rationality and the structure of human memory. *Synthese* 57:163-86.

Furlong, E. J. 1956. The empiricist theory of memory. *Mind* 65:542-47.

Furlong, E. J. 1951. *A Study in Memory: A Philosophical Essay*. Nelson.

Gennaro, R. J. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333-47.

Haight, D. & Haight, M. 1989. Time, memory, and self-remembering. *Journal of Speculative Philosophy* 3:1-11.

Holland, R. F. 1954. The empiricist theory of memory. *Mind* 63:464-86.

Judson, L. 1988. Russell on memory. *Proceedings of the Aristotelian Society* 88:65-82.

Kurtzman, H. S. 1983. Modern conceptions of memory. *Philosophy and Phenomenological Research* 44:1-20.

Malcolm, N. 1970. Memory and representation. *Nous* 4:59-71.

Martin, M. G. F. 1992. Perception, concepts, and memory. *Philosophical Review* 101:745-63.

Munsat, S. 1979. Memory and causality. In (D. Gustafson, ed) *Body, Mind, and Method*. Reidel.

Naylor, A. 1985. In defense of a nontraditional theory of memory. *Monist* 62:136-50.

- Owens, D. 1996. A Lockean theory of memory experience. *Philosophy and Phenomenological Research* 56:319-32.
- Perkins, R. K. 1973. Russell on memory. *Mind* 82:600-1.
- Rakover, S. 1983. In defense of memory viewed as stored mental representation. *Behaviorism* 11:53-62.
- Rosen, D. A. 1975. An argument for the logical notion of a memory trace. *Philosophy of Science* 42:1-10.
- Rundle, B. 1986. Memory and causation. *Philosophical Investigations* 9:302-7.
- Rychlak, J. F. 1996. Memory: A logical learning account. *Journal of Mind and Behavior* 17:229-50.
- Sanders, J. T. 1985. Experience, memory, and intelligence. *Monist* 68:507-21.
- Schectman, M. 1994. The truth about memory. *Philosophical Psychology* 7:3-18.
- Shope, R. K. 1973. Remembering, knowledge, and memory traces. *Philosophy and Phenomenological Research* 33:303-22.
- Stern, D. G. 1991. Models of memory: Wittgenstein and cognitive science. *Philosophical Psychology* 4:203-18.
- Sutton, J. 1998. *Philosophy and Memory Traces: Descartes to Connectionism*. Cambridge University Press.
- Urmson, J. O. 1971. Memory and imagination. *Mind* 80:607.
- Wilcox, S. & Katz, S. 1981. A direct realistic alternative to the traditional conception of memory. *Behaviorism* 9:227-40.
- Zemach, E. 1983. Memory: What it is, and what is cannot possibly be. *Philosophy and Phenomenological Research* 44:31-44.

5.2e Color [see also 1.3a, 1.7a, 1.7d]

- Armstrong, D. M. 1969. Colour realism and the argument from microscopes. In (R. Brown & C. Rollins, eds) *Contemporary Philosophy in Australia*. Humanities Press.
- Averill, E. W. 1985. Color and the anthropocentric problem. *Journal of Philosophy* 82:281-303.

- Averill, E. W. 1992. The relational nature of color. *Philosophical Review* 101:551-88.
- Bigelow, J. Collins, J. & Pargetter, R. 1990. Colouring in the world. *Mind* 99:279-88.
- Boghossian, P. & Velleman, J. D. 1989. Color as a secondary quality. *Mind* 98:81-103.
- Boghossian, P. & Velleman, J. D. 1991. Physicalist theories of color. *Philosophical Review* 100:67-106.
- Broackes, J. 1992. The autonomy of colour. In (D. Charles & K. Lennon, ed) *Reduction, Explanation, and Realism*. Oxford University Press.
- Broackes, J. 1997. *The Nature of Colour*. Routledge.
- Byrne, A. & Hilbert, D. R. 1997. *Readings on Color, Volume 1: The Philosophy of Color*. MIT Press.
- Byrne, A. & Hilbert, D. R. 1997. *Readings on Color, Volume 2: The Science of Color*. MIT Press.
- Byrne, A. & Hilbert, D. R. 1997. Colors and reflectances. In (A. Byrne & D. R. Hilbert, eds) *Readings on Color, Volume 1: The Philosophy of Color*. MIT Press.
- Campbell, J. 1993. A simple view of colour. In (J. Haldane & C. Wright, ed) *Reality, Representation, and Projection*. Oxford University Press.
- Campbell, K. 1969. Colours. In (R. Brown & C. Rollins, eds) *Contemporary Philosophy in Australia*. Humanities Press.
- Campbell, K. 1982. The implications of Land's theory of colour vision. In (L. Cohen, ed) *Logic, Methodology, and Philosophy of Science*, Vol. 6. North-Holland.
- Campbell, K. 1993. David Armstrong and realism about colour. In (J. Bacon, K. Campbell, & L. Reinhardt, eds) *Ontology, Causality, and Mind*. Cambridge University Press.
- Clark, A. 1996. True theories, false colors. *Philosophy of Science Supplement* 63:143-50.
- Dedrick, D. 1995. Objectivism and the evolutionary value of color vision. *Dialogue* 34:35-44.
- Dedrick, D. 1996. Can color be reduced to anything? *Philosophy of Science Supplement* 3:134-42.
- Foti, V. M. 1990. The dimension of color. *International Studies in Philosophy* 22:13-28.

- Gilbert, P. 1987. Westphal and Wittgenstein on white. *Mind* 76:399-403.
- Gilbert, P. 1989. Reflections on white: A rejoinder to Westphal. *Mind* 98:423-6.
- Gold, I. 1999. Dispositions and the central problem of color.
- Gold, I. 1999. On Lewis on naming the colours. *Australasian Journal of Philosophy* 77:365-370.
- Philosophical Studies* 93:21-44.
- Hall, R. J. 1996. The evolution of color vision without colors. *Philosophy of Science Supplement* 63:125-33.
- Hardin, C. L. 1983. Colors, normal observers and standard conditions. *Journal of Philosophy* 80:806-13.
- Hardin, C. L. 1984. A new look at color. *American Philosophical Quarterly* 21:125-33.
- Hardin, C. L. 1984. Are scientific objects colored? *Mind* 93:491-500.
- Hardin, C. L. 1985. The resemblances of colors. *Philosophical Studies* 48:35-47.
- Hardin, C. L. 1985. Frank talk about the colors of sense-data. *Australasian Journal of Philosophy* 63:485-93.
- Hardin, C. L. 1988. *Color for Philosophers*. Hackett.
- Hardin, C. L. 1988. Phenomenal colors and sorites. *Nous* 22:213-34.
- Hardin, C. L. 1989. Could white be green? *Mind* 390:285-8.
- Hardin, C. L. 1989. Idle colors and busy spectra. *Analysis* 49:47-8.
- Hardin, C. L. 1990. Color and illusion. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.
- Hardin, C. L. 1993. van Brakel and the not-so-naked emperor. *British Journal for the Philosophy of Science* 44:137-50.
- Harvey, J. 1992. Challenging the obvious: The logic of color concepts. *Philosophia* 21:277-94.
- Harvey, J. 2000. Colour-dispositionalism and its recent critics. *Philosophy and Phenomenological Research* 61:137-156.

- Hazen, A. P. 1999. On naming the colours. *Australasian Journal of Philosophy* 77:224-231.
- Hilbert, D. R. 1987. *Color and Color Perception: A Study in Anthropocentric Realism*. CSLI Press.
- Hilbert, D. R. 1992. What is color vision? *Philosophical Studies* 68:351-70.
- Jackson, F. 1996. The primary quality view of color. *Philosophical Perspectives* 10:199-219.
- Jackson, F. 1998. Colour, disjunctions, programming. *Analysis* 58:86-88.
- Jackson, F. & Pargetter, R. 1987. An objectivist's guide to subjectivism about color. *Revue Internationale de Philosophie* 41:127-v41.
- Jacovides, M. 2000. Cambridge changes of color. *Pacific Philosophical Quarterly* 81:142-164.
- Johnston, M. 1992. How to speak of the colors. *Philosophical Studies* 68:221-263.
- Kliewer, G. 1998. Neutral color concepts. *Philosophical Studies* 91:21-41.
- Kraut, R. 1992. The objectivity of color and the color of objectivity. *Philosophical Studies* 3:265-87.
- Langsam, H. 2000. Why colours do look like dispositions. *Philosophical Quarterly* 50:68-75.
- Landesman, C. 1989. *Color and Consciousness: An Essay in Metaphysics*. Temple University Press.
- Levin, J. 2000. Dispositional theories of color and the claims of common sense. *Philosophical Studies* 100:151-174.
- Lewis, D. 1997. Naming the colours. *Australasian Journal of Philosophy* 75:325-42.
- Maund, J. B. 1981. Colour: A case for conceptual fission. *Australasian Journal of Philosophy* 59:308-22.
- Maund, J. B. 1991. The nature of color. *History of Philosophy Quarterly* 8:253-63.
- Maund, J. B. 1995. *Colours: Their Nature and Representation*. Cambridge University Press.
- McFarland, D. & Miller, A. 1998. Jackson on colour as a primary quality. *Analysis* 58:76-85.
- McFarland, D. & Miller, A. 2000. Disjunctions, programming and the Australian view of colour. *Analysis* 60:209-212.

- McGilvray, J. A. 1983. To color. *Synthese* 54:37-70.
- McGilvray, J. A. 1994. Constant colors in the head. *Synthese* 100:197-239.
- McGinn, C. 1996. Another look at color. *Journal of Philosophy* 93:537-53.
- McGinn, M. 1991. Westphal on the physical basis of color incompatibility. *Analysis* 4:218-22.
- McGinn, M. 1991. On two recent accounts of color. *Philosophical Quarterly* 41:316-24.
- Miller, A. 2001. The missing-explanation argument revisited. *Analysis* 61:76-86.
- Montgomery, R. 1996. The indeterminacy of color vision. *Synthese* 106:167-203.
- Nida-Rumelin, M. 1997. The character of color predicates: A phenomenalist view. In (M. Anduscius, A. Newen, & W. Kunne, eds) *Direct Reference, Indexicality, and Propositional Attitudes*. CSLI Press.
- Ross, P. W. 1999. The appearance and nature of color. *Southern Journal of Philosophy* 37:227-252.
- Ross, P. 2000. The relativity of color. *Synthese* 123:105-130.
- Smart, J. J. C. 1975. On some criticisms of a physicalist theory of colors. In (C. Cheng, ed) *Philosophical Aspects of the Mind-Body Problem*. University Press of Hawaii.
- Smart, J. J. C. 1995. 'Looks red' and dangerous talk. *Philosophy* 70:545-54.
- Smith, M. A. Color, transparency, mind-independence. In (J. Haldane & C. Wright, ed) *Reality, Representation, and Projection*. Oxford University Press.
- Smith, P. 1987. Subjectivity and colour vision. *Proceedings of the Aristotelian Society* 61:245-81.
- Spohn, W. 1997. The character of color predicates: A materialist view. In (M. Anduscius, A. Newen, & W. Kunne, eds) *Direct Reference, Indexicality, and Propositional Attitudes*. CSLI Press.
- Strawson, G. 1989. Red and 'red'. *Synthese* 78:193-232.
- Stroud, B. 2000. *The Quest for Reality: Subjectivism and the Metaphysics of Colour*. Oxford University Press.
- Stroud-Drinkwater, C. 1994. The naive theory of color. *Philosophy and Phenomenological Research*

54:345-54.

Thompson, E. , Palacios, A. , & Varela, F. J. 1992. Ways of coloring. Behavioral and Brain Sciences.

Thompson, E. 1995. Colour vision, evolution, and perceptual content. Synthese 104:1-32.

Thompson, E. 1995. *Colour Vision*. Routledge.

Tolliver, J. T. 1996. Interior colors. Philosophical Topics 22:411-41.

van Brakel, J. 1993. The plasticity of categories: The case of color. British Journal for the Philosophy of Science 44:103-135.

Watkins, M. 1999. Do animals see colors? An anthropocentrist's guide to animals, the color blind, and far away places. Philosophical Studies 94:189-209.

Westphal, J. 1982. Brown: Remarks on color. Inquiry 25:417-33.

Westphal, J. 1986. White. Mind 95:310-28.

Westphal, J. 1989. Black. Mind 98:585-9.

Westphal, J. 1991. *Colour: A Philosophical Introduction*. Blackwell.

Whitmyer, VG. 1999. Ecological color. Philosophical Psychology 12:197-214.

Wittgenstein, L. 1977. *Remarks on Colour*. University of California Press.

5.3 Philosophy of Psychology, General

5.3a

Psychological Laws [see also 3.5d]

Antony, L. 1995. Law and order in psychology. Philosophical Perspectives 9:429-46.

Braithwaite, M. 1949. Causal laws in psychology. Aristotelian Society Supplement 23:45-60.

Fodor, J. A. 1991. You can fool some of the people all of the time, everything else being equal: Hedged laws and psychological explanation. Mind 100:19-34.

Ceteris paribus means that every realizing state has completing conditions. Even absolute exceptions are OK, as long as they're not across-the-board.

Fodor, J. A. 1989. Making mind matter more. *Philosophical Topics* 17:59-79. Reprinted in *A Theory of Content and Other Essays* (MIT Press, 1990).

Non-strict psychological laws are compatible with the (nomologically sufficient) causal responsibility of mental properties. So there's no need for epiphobia. With comments on the relation between laws and mechanisms.

Horgan, T. & Tienson, J. 1990. Soft laws. *Midwest Studies in Philosophy* 15.

Argues that any laws in intentional psychology have ineliminable same-level exceptions; the Kuhnian crisis in cognitive science gives evidence for this. But ceteris paribus laws provide perfectly good theoretical explanation.

Lycan, W. G. 1981. Psychological laws. *Philosophical Topics* 12:9-38.

A functionalist defense against anomalous monism. Psychofunctional laws and psychological laws, though not psychophysical laws, may exist. Rebutting arguments from rationality, indeterminism, intensionality, etc.

Mace, C. A. 1949. Causal laws in psychology. *Aristotelian Society Supplement* 23:61-68.

Marcello, G. 2000. Horgan and Tienson on ceteris paribus laws. *Philosophy of Science* 67:301-315.

Mott, P. 1992. Fodor and ceteris paribus laws. *Mind* 101:335-46.

Pietroski, P. & Rey, G. 1995. When other things aren't equal: Saving ceteris paribus laws from vacuity. *British Journal for the Philosophy of Science* 46:81-110.

Schiffer, S. 1991. Ceteris paribus laws. *Mind* 100:1-17.

There are no ceteris paribus laws, as there's no satisfactory way to cash the "unless" cause. But psychology doesn't need laws, anyway.

Silverberg, A. 1996. Psychological laws and nonmonotonic logic. *Erkenntnis* 44:199-224.

Warfield, T. A. 1993. Folk-psychological ceteris-paribus laws. *Philosophical Studies* 71:99-112.

5.3b Psychology and Neuroscience [see also 6.1i]

Bechtel, W. 1983. A bridge between cognitive science and neuroscience: The functional architecture of mind. *Philosophical Studies* 44:319-30.

Arguing for the notion of functional architecture as a bridge whereby neural components can be components of cognitive processes.

Bub, J. 1994. Testing models of cognition through the analysis of brain-damaged patients. *British Journal for the Philosophy of Science* 45:837-55.

Butler, K. 1994. Neural constraints in cognitive science. *Minds and Machines* 4:129-62.

Cherniak, C. 1991. Meta-neuroanatomy: The myth of the unbounded mind/brain. In (E. Agazzi, ed) *Philosophy and the Origin and Evolution of the Universe*.

Cherniak, C. 1994. Philosophy and computational neuroanatomy. *Philosophical Studies* 73:89-107.

Argues that we can understand the brain under the hypothesis that it is optimized to "save wire", due to bounded resources: organization predicts placement. With remarks on the relation between cognitive and neural levels.

Churchland, P. M. 1986. Some reductive strategies in cognitive neurobiology. *Mind* 95:279-309. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Some cute examples of neurophysiological reductions using state-spaces.

Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. MIT Press.

Churchland, P. S. 1980. A perspective on mind-brain research. *Journal of Philosophy* 77:185-207.

The brain can tell us a lot about the mind. With examples.

Churchland, P. S. 1982. Mind-brain reduction: New light from philosophy of science. *Neuroscience* 7:1041-7.

Churchland, P. S. 1986. *Neurophilosophy: Toward A Unified Science of the Mind-Brain*. MIT Press.

All about neuroscience, philosophy and prospects for their interaction.

Churchland, P. S. & Sejnowski, T. 1989. Neural representation and neural computation. In (L. Nadel, ed) *Neural Connections, Mental Computations*. MIT Press.

About how neuroscience and connectionism affect our conception of mind.

Churchland, P. S. 1987. Epistemology in the age of neuroscience. *Journal of Philosophy* 84:546-53.

On paradigm shifts, biology, evolution, connectionism, etc.

Clark, A. 1980. *Psychological Models and Neural Mechanisms: An Examination of Reductionism in Psychology*. Oxford University Press.

Glymour, C. 1994. On the methods of cognitive neuropsychology. *British Journal for the Philosophy of Science* 45:815-35.

Hardcastle, V. G. 1992. Reduction, explanatory extension, and the mind/brain sciences. *Philosophy of Science* 59:408-28.

The relationship between psychology and neuroscience is best characterized not by reduction but by explanatory extension, where each field is enriched by the other. With a number of examples from recent empirical work.

Hatfield, G. 1988. Neurophilosophy meets psychology: Reduction, autonomy, and empirical constraints. *Cognitive Neuropsychology* 5:723-46.

Hatfield, G. 2000. The brain's 'new' science: Psychology, neurophysiology, and constraint. *Philosophy of Science* 67:388-404.

Klagge, J. C. 1989. Wittgenstein and neuroscience. *Synthese* 78:319-43.

Wittgenstein wouldn't have liked the Churchlands, as neuro might be chaos, and too much neuro might undermine our self-conception nihilistically.

Kobes, B. 1991. On a model for psycho-neural coevolution. *Behavior and Philosophy* 19:1-17.

Madell, G. 1986. Neurophilosophy: A principled skeptic's response. *Inquiry*.

Manier, E. 1986. Problems in the development of cognitive neuroscience: Effective communication between scientific domains. *Philosophy of Science Association* 1986, 1:183-97.

McCauley, R. 1986. Intertheoretic relations and the future of psychology. *Philosophy of Science* 53:179-99.

Incommensurable theories don't necessarily require elimination, if their relationship is synchronic/interlevel, rather than diachronic/intralevel.

Mucciolo, L. 1974. The identity thesis and neuropsychology. *Nous* 8:327-42.

Argues contra Fodor and Block that neurological equipotentiality doesn't refute type materialism. Mental states may not be anatomically defined neural states, but they may be more abstract neural holograms.

Mundale, J. & Bechtel, W. 1996. Integrating neuroscience, psychology, and evolutionary biology through a teleological conception of function. *Minds and Machines* 6:481-505.

Ravenscroft, I. 1998. Neuroscience and the mind. *Mind and Language* 13:132-137.

Rockwell, W. T. 1994. On what the mind is identical with. *Philosophical Psychology* 7:307-23.

Argues that the mind is not identical with the brain -- at the very least, it's the central nervous system, and perhaps more. "Brain" does not denote a natural kind in neurophysiology.

Smith, A. 1986. Brain-mind philosophy. *Inquiry* 29:203-15.

Skarda, S. 1986. Explaining behavior: Bringing the brain back in. *Inquiry* 29:187-201.

Stoljar, D. & Gold, S. 1998. On biological and cognitive neuroscience. *Mind and Language* 13:110-31.

Stone, T. & Davies, M. 1993. Cognitive neuropsychology and the philosophy of mind. *British Journal for the Philosophy of Science* 44:589-622.

Stufflebeam, R. S. & Bechtel, W. 1997. PET: Exploring the myth and the method. *Philosophy of Science* 64:95-106.

van Orden, G. C. 1997. Functional neuroimages fail to discover pieces of mind in the parts of the brain. *Philosophy of Science Supplement* 64:85-94.

von Eckardt, B. 1984. Cognitive psychology and principled skepticism. *Journal of Philosophy* 81:67-88.

Cognitive psychology can transmogrify itself, who needs neuroscience?

5.3c Explanation in Cognitive Science

Cummins, R. 1982. The internal manual model of psychological explanation. *Cognition and Brain Theory* 5:257-68.

Cummins, R. 1983. *The Nature of Psychological Explanation*. MIT Press.

Psychological explanation is typically via functional analysis, not causal subsumption. On interpretation, computation, and an analysis of cognition and intentionality. With remarks on Dretske, Searle, Titchener, Hull, Freud.

Fodor, J. A. 1968. *Psychological Explanation*. Random House.

Fodor, J. A. 1968. The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy* 65:627-40. Reprinted in *RePresentations* (MIT Press, 1980).

Franks, B. 1995. On explanation in cognitive science: Competence, idealization, and the failure of the classical cascade. *British Journal for the Philosophy of Science* 46:475-502.

Gilman, D. 1993. Optimization and simplicity: Marr's theory of vision and biological explanation. *Synthese* 107:293-323.

Contra Kitcher 1988, much of Marr's theory doesn't depend on optimization; in any case, optimization isn't so bad. With remarks on interdisciplinarity.

Heil, J. 1986. Formalism and psychological explanation. *Journal of Mind and Behavior* 7:1-10.

On the tension between formal explanation and representational explanation.

Kim, J. 1989. Mechanism, purpose, and explanatory exclusion. *Philosophical Perspectives* 3:77-108. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Discusses the principle: there cannot be two independent explanations of the same phenomena. With application to purposive explanation of behavior, theory reduction, and eliminativism, and a discussion of explanatory realism.

Kim, J. 1990. Explanatory exclusion and the problem of mental causation. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

On the problems posed by explanatory exclusion, and possible solutions. With focus on the problems as they arise for Dretske's and Davidson's theories.

Knight, D. 1997. A poetics of psychological explanation. *Metaphilosophy* 28:63-80.

Millikan, R. G. 1993. Explanation in biopsychology. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Montgomery, R. 1995. Explanation and evaluation in cognitive science. *Philosophy of Science* 62:261-82.

- Montgomery, R. 1998. Grades of explanation in cognitive science. *Synthese* 114:463-495.
- Morris, M. 1986. Causes of behavior. *Philosophical Quarterly* 36:123-44.
- Moser, P. 1994. Naturalism and psychological explanation. *Philosophical Psychology* 7:63-84.
- Owens, J. 1998. Psychological explanation and causal deviancy. *Synthese* 115:143-169.
- Sober, E. 1978. Psychologism. *Journal for the Theory of Social Behavior* 8:165-91.

5.3d

Philosophy of Cognitive Science, Misc

- Bealer, G. 1987. The boundary between philosophy and cognitive science. *Journal of Philosophy* 86:553-55.
- Philosophy is autonomous: empirical considerations can't affect it.
- Bogdan, R. 2000. *Minding Minds: Evolving a Reflexive Mind by Interpreting Others*. MIT Press.
- Fetzer, J. H. 1991. *Philosophy and Cognitive Science*. Paragon House.
- Flanagan, O. J. 1984. *The Science of the Mind*. MIT Press.
- Hardcastle, V. 1996. *How to Build a Theory in Cognitive Science*. SUNY Press.
- Harnad, S. 1982. Neoconstructivism: A unifying constraint for cognitive science. In (T. Simon & R. Scholes, eds) *Language, Mind, and Brain*. Lawrence Erlbaum.
- Haugeland, J. 1978. The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* 1:215-26.
- Hooker, C. A. 1975. The information-processing approach to the brain-mind and its philosophical ramifications. *Philosophy and Phenomenological Research* 36:1-15.
- Keely, B. 2000. Neuroethology and the philosophy of cognitive science. *Philosophy of Science* 67:404-418.

- Kukla, A. 1989. Non-empirical issues in psychology. *American Psychologist* 44:485-94.

On the role of non-empirical advances in psychology: e.g. in theory construction, coherence analysis,

conceptual innovation, with the aid of logically necessary truths and the contingent/pragmatic a priori.

Lloyd, D. 1989. *Simple Minds*. MIT Press.

O'Nuillain, S. 1995. *The Search for Mind: A New Foundation for Cognitive Science*. Ablex.

O'Nuillain, S. , McKevitt, P. & MacAogain, E. (eds) 1997. *Two Sciences of Mind*. John Benjamins.

Pickering, M. & Chater, N. 1995. Why cognitive science is not formalized folk psychology. *Minds and Machines* 5.

Preston, B. 1994. Behaviorism and mentalism: Is there a third alternative? *Synthese* 100:167-96.

van Gelder, T. 1998. The roles of philosophy in cognitive science. *Philosophical Psychology* 11:117-36.

von Eckardt, B. 1993. *What is Cognitive Science?* MIT Press.

5.4

Philosophy of Mind, General

Armstrong, D. 1999. *The Mind-Body Problem: An Opinionated Introduction*. Westview Press.

Baker, L. R. 1989. Recent work in the philosophy of mind. *Philosophical Books* 30:1-9.

A general overview.

Bealer, G. 1986. The logical status of mind. *Midwest Studies in Philosophy* 10.

Bechtel, W. 1988. *Philosophy of Mind: An Overview for Cognitive Science*. Lawrence Erlbaum.

Braddon-Mitchell, D. & Jackson, F. 1997. *Philosophy of Mind and Cognition*. Blackwell.

Burge, T. 1992. Philosophy of language and mind: 1950-1990. *Philosophical Review* 100:3-52.

An overview of the last 40 years of the philosophy of language and the philosophy of mind, covering many issues and trends.

Carruthers, P. 1986. *Introducing Persons: Theories and Arguments in the Philosophy of Mind*. SUNY Press.

Churchland, P. M. 1984. *Matter and Consciousness*. MIT Press.

Crane, T. 1995. *The Mechanical Mind*. Penguin.

Dennett, D. C. 1978. Current issues in the philosophy of mind. *American Philosophical Quarterly* 15:249-261.

An overview of everything, circa 1978: logical behaviorism, functionalism, the identity theory, qualia, meaning, and so on, with bibliography.

Graham, G. 1993. *Philosophy of Mind: An Introduction*. Blackwell.

Guttenplan, S. 2000. *Mind's Landscape: An Introduction to the Philosophy of Mind*. Blackwell Publishers.

Haldane, J. J. 1994. Analytical philosophy and the nature of mind: Time for another rebirth? In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Haldane, J. 2000. The state and fate of contemporary philosophy of mind. *American Philosophical Quarterly* 37:301-21.

Hannay, B. 1994. *Subjectivity and Reduction: An Introduction to the Mind-Body Problem*. Westview Press.

Harman, G. 1989. Some philosophical issues in cognitive science. In (M. Posner, ed) *Foundations of Cognitive Science*. MIT Press.

Kim, J. 1996. *Philosophy of Mind*. Westview Press.

Lowe, E. J. 2000. *An Introduction to the Philosophy of Mind*. Cambridge University Press.

McGinn, C. 1982. *The Character of Mind*. Oxford University Press.

Phillips, H. 1995. *Vicissitudes of the I: An Introduction to the Philosophy of Mind*. Prentice-Hall.

Quine, W. V. 1985. States of mind. *Journal of Philosophy* 82:5-8.

Rey, G. 1997. *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Blackwell.

Rorty, R. 1982. Contemporary philosophy of mind. *Synthese* 53:323-48.

In praise of the "Ryle-Dennett" tradition, and the elimination of dualism from the philosophy of mind.

Rorty, R. 1993. Consciousness, intentionality, and pragmatism. In (S. Christensen & D. Turner, eds) *Folk Psychology and the Philosophy of Mind*. Lawrence Erlbaum.

A pragmatist perspective on the recent history of the philosophy of mind, focusing on consciousness, intentionality, and mental representation, and on debates between Fodor, Dennett, Searle, Putnam, and Davidson.

Shaffer, J. A. 1964. *Philosophy of Mind*. Prentice-Hall.

Smith, P. & Jones, O. 1986. *The Philosophy of Mind: An Introduction*. Cambridge University Press.

Sprague, E. 1999. *Persons and their Minds: A Philosophical Investigation*. Westview Press.

Part 6: Consciousness in Science [1599]

Part of [Contemporary Philosophy of Mind: An Annotated Bibliography](#).

Compiled by [David J. Chalmers](#), Department of Philosophy, University of Arizona, Tucson AZ 85721. E-mail: chalmers@arizona.edu.

- 6.1 [Consciousness and Neuroscience](#) [460]
 - 6.1a [Neural Correlates of Consciousness](#) [74]
 - 6.1b [Neural Correlates of Visual Consciousness](#) [28]
 - 6.1c [Blindsight](#) [50]
 - 6.1d [Other Neuropsychological Disorders](#) [71]
 - 6.1e [Cerebral Hemispheres and Consciousness](#) [45]
 - 6.1f [Neural Timing and Consciousness](#) (Libet) [19]
 - 6.1g [Neural Synchrony and Binding](#) [42]
 - 6.1h [Consciousness and Anesthesia](#) [44]
 - 6.1i [Foundational Issues](#) [31]
 - 6.1j [Consciousness and Neuroscience, Misc](#) [56]
- 6.2 [Consciousness and Psychology](#) [787]
 - 6.2a [Cognitive Models of Consciousness](#) [60]
 - 6.2b [Unconscious Perception](#) [73]
 - 6.2c [Unconscious Processes](#) [45]
 - 6.2d [Visual Consciousness](#) [37]
 - 6.2e [Consciousness and Attention](#) [35]
 - 6.2f [Consciousness and Memory](#) [84]
 - 6.2g [Consciousness and Learning](#) [33]
 - 6.2h [Consciousness and Metacognition](#) [22]
 - 6.2i [Consciousness and Control](#) [38]
 - 6.2j [Consciousness and Imagery](#) [24]
 - 6.2k [Consciousness and Emotion](#) [31]
 - 6.2l [Consciousness, Sleep, and Dreaming](#) [34]
 - 6.2m [Consciousness in Psychoanalysis](#) [57]
 - 6.2n [Consciousness and Time](#) [35]
 - 6.2o [Self-Consciousness](#) [46]
 - 6.2p [Development of Consciousness](#) [33]
 - 6.2q [The Stream of Consciousness](#) [29]
 - 6.2r [Foundational Issues](#) [19]
 - 6.2s [Consciousness and Psychology, Misc](#) [52]

- 6.3 [Consciousness and Physics](#) [77]
 - 6.3a [Interpretation of Quantum Mechanics](#) [27]
 - 6.3b [Quantum Mechanisms of Consciousness](#) [30]
 - 6.3c [Consciousness and Physics, Misc](#) [20]
- 6.4 [Consciousness and Science, Misc](#) [275]
 - 6.4a [Evolution of Consciousness](#) [31]
 - 6.4b [Consciousness and Language](#) [23]
 - 6.4c [Animal Consciousness](#) [33]
 - 6.4d [Animal Self-Consciousness](#) [22]
 - 6.4e [Altered States of Consciousness](#) [27]
 - 6.4f [Parapsychology and Consciousness](#) [54]
 - 6.4g [Phenomenology](#) [34]
 - 6.4h [Foundations](#) [27]
 - 6.4i [Miscellaneous](#) [24]

6.1 Consciousness and Neuroscience

6.1a

Neural Correlates of Consciousness

Alkire, M. T. , Haier, R. J. , & James, H. F. 1998. Toward the neurobiology of consciousness: Using brain imaging and anesthesia to investigate the anatomy of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Baars, B. J. & Newman, J. 1994. A neurobiological interpretation of the Global Workspace theory of consciousness. In (A. Revonsuo & M. Kamppinen, eds) *Consciousness in Philosophy and Cognitive Neuroscience*. Lawrence Erlbaum.

Baars, B. J. 1995. Surprisingly small subcortical structures are needed for the state of waking consciousness, while cortical projection areas seem to provide perceptual contents of consciousness. *Consciousness and Cognition* 4:159-62.

Baars, B. J. , Newman, J. & Taylor, J. G. 1998. Neuronal mechanisms of consciousness: A relational global workspace approach. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Bogen, J. E. 1995. On the neurophysiology of consciousness, part I: An overview. *Consciousness and Cognition* 4:52-62.

- Bogen, J. E. 1995. On the neurophysiology of consciousness, part II: Constraining the semantic problem. *Consciousness and Cognition* 4:137-58.
- Bogen, J. E. 1997. Some neurophysiologic aspects of consciousness. *Seminars in Neurology* 17:95-103.
- Bogen, J. E. 1998. Locating the subjectivity pump: The thalamic intralaminar nuclei. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Bremer, F. 1966. Neurophysiological correlates of mental unity. In (J. Eccles, ed) *Brain and Conscious Experience*. Springer.
- Buser, P. A. & Rougeul-Buser, A. (eds) 1978. Cerebral correlates of conscious experience. Elsevier.
- Coenen, A. M. L. 1998. Neuronal phenomena associated with vigilance and consciousness: From cellular mechanisms to electroencephalographic patterns. *Consciousness and Cognition* 7:42-53.
- Crick, F. 1984. Functions of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences USA* 81:4586-93.
- Crick, F. & Koch, C. 1990. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2:263-275.
- Crick, F. & Koch, C. 1998. Consciousness and neuroscience. *Cerebral Cortex*.
- Crick, F. & Koch, C. 2000. The unconscious homunculus. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Damasio, A. 2000. A neurobiology for consciousness. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Dimond, S. 1976. Brain circuits for consciousness. *Brain, Behavior, and Evolution* 13:376-95.
- Duzel E. , Yonelinas A. P. , Mangun G. R. , Heinze H. J. , & Tulving E. 1997. Event-related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences of the United States of America* 94:5973-8.
- Edelman, G. & Tononi, G. 2000. Reentry and the dynamic core: Neural correlates of conscious experience. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Ellis, R. 2000. Efferent brain processes and the enactive approach to consciousness. *Journal Of Consciousness Studies* 7:40-50.

- Flohr, H. 1990. Brain processes and phenomenal consciousness: A new and specific hypothesis. *Theory and Psychology* 1:245-62.
- Flohr, H. 1992. Qualia and brain processes. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.
- Flohr, H. 1995. Sensations and brain processes. *Behavioral Brain Research* 71:157-61.
- Gallese, V. 2000. The acting subject: Toward the neural basis of social cognition. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Gazzaniga, M. 1993. Brain mechanisms and conscious experience. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.
- Gazzaniga, M. 1998. Brain and conscious experience. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.
- Gray, J. A. 1995. The contents of consciousness: A neuropsychological conjecture. *Behavioral and Brain Sciences* 18:659-76.
- Greenfield, S. 1997. How might the brain generate consciousness? *Communication and Cognition* 30:285-300.
- Greenfield, S. 1998. A rosetta stone for mind and brain? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Grossenbacher, P. G. (ed) 1997. *Finding Consciousness in the Brain: A Neurocognitive Approach*. John Benjamins.
- Goldman-Rakic, P. S. 1988. The prefrontal contribution to working memory and conscious experience. In (O. Creutzfeld & J. Eccles, eds) *The Brain and Conscious Experience*. Pontifical Academy.
- Hobson, J. A. 1994. *The Chemistry of Conscious States*. Basic Books.
- Hobson, J. A. 1997. Consciousness as a state-dependent phenomenon. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Jasper, H. 1998. Sensory information and conscious experience. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.
- John, E. R. , Easton, P. & Isenhardt, R. 1997. Consciousness and cognition may be mediated by multiple

independent coherent ensembles. *Consciousness and Cognition* 6:3-39.

Jones, B. E. 1998. The neural basis of consciousness across the sleep-waking cycle. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Jung, R. 1954. Correlation of bioelectrical and autonomic phenomena with alterations of consciousness and arousal in man. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.

Kahn, D. , Pace-Schott, E. F. & Hobson, J. A. 1997. Consciousness in waking and dreaming: The roles of neuronal oscillation and neuromodulation in determining similarities and differences. *Neuroscience* 78:13-38.

Kanwisher, N. 2001. Neural events and perceptual awareness. *Cognition* 79:89-113.

Kinsbourne, M. 1988. An integrated field theory of consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Kinsbourne, M. 1993. Integrated cortical field model of consciousness. In (Ciba Foundation) *Experimental and Theoretical Studies of Consciousness*. Wiley.

Kinsbourne, M. 1995. The intralaminar thalamic nuclei: Subjectivity pumps or attention-action coordinators? *Consciousness and Cognition* 4:167-71.

Kleitman, N. 1955. The role of the cerebral cortex in the development and maintenance of consciousness. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Third Conference*. Josiah Macy Foundation.

Koch, C. & Crick, F. 1994. Some further ideas regarding the neuronal basis of awareness. In (C. Koch & J. Davis, eds) *Large-Scale Neuronal Theories of the Brain*. MIT Press.

Koch, C. & Crick, F. 2000. Some thoughts on consciousness and neuroscience. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.

Lehmann, D. , Strik, W. K. , Henggeler, B. , Koenig, T. 1998. Brain electric microstates and momentary conscious mind states as building blocks of spontaneous thinking: I. Visual imagery and abstract thoughts. *International Journal of Psychophysiology* 29:1-11.

Libet, B. 1982. Brain stimulation in the study of neuronal functions for conscious sensory experiences. *Human Neurobiology* 1:235-42.

Libet, B. 1989. Conscious subjective experience vs. unconscious mental functions: A theory of the

cerebral processes involved. In (R. Cotterill, ed) *Models of Brain Function*. Cambridge University Press.

Libet, B. 1996. Neural processes in the production of conscious experiences. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Libet, B. 1998. Do the models offer testable proposals of brain functions for conscious experience? In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Magoun, H. W. 1954. The ascending reticular system and wakefulness. In (J. Delafresnaye, ed) *Brain Mechanism and Consciousness*. Blackwell.

Markowitsch H. J. 1995. Cerebral bases of consciousness: A historical view. *Neuropsychologia* 33:1181-1192.

Metzinger, T. 2000. *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. MIT Press.

Newman, J. B. 1995. Thalamic contributions to attention and consciousness. *Consciousness and Cognition* 4:172-93.

Newman, J. B. 1997. Putting the puzzle together: Toward a general theory of the neural correlates of consciousness. *Journal of Consciousness Studies* 4:47-66, 4:100-121.

Newman, J. B. & Baars, B. J. 1993. A neural attentional model for access to consciousness: A global workspace perspective. *Concepts in Neuroscience* 4:255-90.

O'Keefe, J. 1985. Is consciousness the gateway to the hippocampal cognitive map? A speculative essay on the neural basis of mind. In (D. Oakley, ed) *Brain and Mind*. Methuen.

Orpwood R. D. 1994. A possible neural mechanism underlying consciousness based on the pattern processing capabilities of pyramidal neurons in the cerebral cortex. *Journal of Theoretical Biology* 169:403-18.

Parvizi, J. & Damasio, A. 2001. Consciousness and the brainstem. *Cognition* 79:135-59.

Penfield, W. 1937. The cerebral cortex and consciousness. In *The Harvey Lectures*. Reprinted in (R. Wilkins, ed) *Neurosurgical Classics*. Johnson Reprint Corporation, 1965.

Penfield, W. 1955. The permanent record of the stream of consciousness. *Acta Psychologica* 11:47-69.

Raichle, M. 2000. The neural correlates of consciousness: An analysis of cognitive skill learning. In (M.

Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.

Rapcsak, S. & Kaszniak, A. 2000. Searching for the neural correlates of consciousness: Clues from face recognition research. *Brain & Cognition* 42:37-40.

Rudell, A. P. & Hua, J. 1996. The recognition potential and conscious awareness. *Electroencephalography and Clinical Neurophysiology* 98:309-318.

Sewards, T. & Sewards, M. 2000. The awareness of thirst: Proposed neural correlates. *Consciousness & Cognition* 9:463-487.

Smythies, J. 1997. The functional neuroanatomy of awareness: With a focus on the role of various anatomical systems in the control of intermodal attention. *Consciousness and Cognition* 6:455-81.

Sokolov, E. N. 1992. The neurophysiological mechanisms of consciousness. *Journal of Russian and East European Psychology* 30:6-12.

Strehler, B. L. 1991. Where is the self?: A neuroanatomical theory of consciousness. *Synapse* 7:44-91.

Stuss, D. T. 1991. Self, awareness, and the frontal lobes: A neuropsychological perspective. In (J. Strauss, ed) *The Self: Interdisciplinary Approaches*. Springer-Verlag.

Taylor, J. G. 1998. Cortical activity and the explanatory gap. *Consciousness and Cognition* 7:109-48.

Tononi, G. & Edelman, G. 1998. Consciousness and the integration of information in the brain. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Umla, C. 2000. Conscious experience depends on multiple brain systems. *European Psychologist* 5:3-11.

Vanderwolf, C. 2000. Are neocortical gamma waves related to consciousness? *Brain Research* 855:217-224.

Verfaellie, M. & Keane, M. M. 1997. The neural basis of aware and unaware forms of memory. *Seminars in Neurology* 17:153-61.

Woolf, N. J. 1997. A possible role for cholinergic neurons of the basal forebrain and pontomesencephalon in consciousness. *Consciousness and Cognition* 6:574-596.

6.1b

Neural Correlates of Visual Consciousness (see also 6.1c)

Anderson, R. A. 1997. Neural mechanisms in visual motion perception in primates. *Neuron* 18:865-872.

Cowey, A. 1996. Visual awareness: Still at sea with seeing? *Current Biology* 6:45-47.

Crick, F. & Koch, C. 1995. Cortical areas in visual awareness. *Nature* 377:294-5.

Crick, F. & Koch, C. 1995. Are we aware of neural activity in primary visual cortex? *Nature* 375:121-23.

Farah, M. J. , O'Reilly, R. C. & Vecera, S. P. 1997. The neural correlates of perceptual awareness: Evidence from covert recognition in prosopagnosia. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Farah, M. 2000. *The Cognitive Neuroscience of Vision*. Blackwell Publishers.

Ffytche, D. 2000. Imaging conscious vision. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Goodale, M. A. & Milner, A. D. 1992. Separate visual pathways for perception and action. *Trends in Neuroscience* 15:20-25.

Goodale, M. & Murphy, K. 2000. Space in the brain: Different neural substrates for allocentric and egocentric frames of reference. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Hubel, D. H. 1998. Recordings from the striate cortex in awake behaving animals. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Koch, C. 1995. Visual awareness and the thalamic intralaminar nuclei. *Consciousness and Cognition* 4:163-66.

Koch, C. 1996. Toward the neuronal substrate of visual consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Koch, C. 1998. The neuroanatomy of visual consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Koch, C. & Braun, J. 1996. Toward the neuronal correlate of visual awareness. *Current Opinion in Neurobiology* 6:158-64.

- Leopold, D. A. & Logothetis, N. K. 1996. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* 379: 549-553.
- Logothetis, N. & Schall, J. 1989. Neuronal correlates of subjective visual perception. *Science* 245:761-63.
- Logothetis, N. K. , Leopold, D. A. & Sheinberg, D. L. 1996. What is rivalling during binocular rivalry? *Nature* 30(6575):621-624.
- Logothetis, N. K. & Leopold, D. A. 1998. Single-neuron activity and visual perception. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Logothetis, N. K. 1999. Binocular rivalry: A window onto consciousness. *Scientific American*.
- Lumer, E. 2000. Binocular rivalry and human visual awareness. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Marzi, C. , Girelli, M. , Miniussi, C. , Smania, N. , & Maravita, A. 2000. Electrophysiological correlates of conscious vision: Evidence from unilateral extinction. *Journal of Cognitive Neuroscience* 12:869-877.
- Milner, A. D. 1995. Cerebral correlates of visual awareness. *Neuropsychologia* 33:1117-30.
- Milner, A. D. 1998. Streams and consciousness: Visual awareness and the brain. *Trends in Cognitive Sciences* 2:25-30.
- Milner, A. D. & Goodale, M. A. 1995. *The Visual Brain in Action*. Oxford University Press.
- Sewards, T. & Sewards, M. 2000. Visual awareness due to neuronal activities in subcortical structures: A proposal. *Consciousness & Cognition* 9:86-116
- Sheinberg, D. L. & Logothetis, N. K. 1997. The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences USA* 94:3408-3413.
- Stoerig P. & Cowey A. 1995. Visual perception and phenomenal consciousness. *Behavioural Brain Research* 71:147-156.
- Vanni, S. , Revonsuo, A. , Saarinen, J. & Hari, R. 1996. Visual awareness of objects correlates with activity of right occipital cortex. *Neuroreport* 8:183-186.
- Zeki, S. & Bartels, A. 1999. Toward a theory of visual consciousness. *Consciousness and Cognition* 8:225-59.

6.1c Blindsight

Azzopardi, P. , & Cowey, A. 1997. Is blindsight like normal, near-threshold vision? *Proceedings of the National Academy of Sciences USA* 94:14190.

Barbur, J. L. , Watson, J. D. G. , Frackowiak, R. D. G. , & Zeki, S. 1993. Conscious visual perception without V1. *Brain* 116:1293-1302.

Braddick, O. , Atkinson, J. , Hood, B. , Harkness, W. 1992. Possible blindsight in infants lacking one cerebral hemisphere. *Nature* 360:461-463.

Campion, J, Latto, R. , & Smith, Y. 1983. Is blindsight an effect of scattered light, spared cortex, and near-threshold vision? *Behavioral and Brain Sciences* 6:423-86.

Carey, D. P. , Goodale, M. A. & Sprowl, E. G. 1990. Blindsight in rodents: The use of a "high-level" distance cue in gerbils with lesions of primary visual cortex. *Behavioural Brain Research* 38:283-289.

Cowey, A. 1995. Blindsight in real sight. *Nature* 377:290-1.

Cowey, A. 1995. Blindsight in monkeys. *Nature* 373:247-9.

Cowey, A. & Stoerig, P. 1991. The neurobiology of blindsight. *Trends in Neurosciences* 14:140-5.

Cowey, A. & Stoerig, P. 1992. Reflections on blindsight. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Cowey, A. & Stoerig, P. 1997. Visual detection in monkeys with blindsight. *Neuropsychologia* 35:929-39.

Danckert, J. & Goodale, M. 2000. Blindsight: A conscious route to unconscious vision. *Current Biology* 10:R64-R67.

Gazzaniga, M. S. , Fendrich, R. & Wessinger, C. M. 1994. Blindsight reconsidered. *Current Directions in Psychological Science* 3:93-96.

Graves, R. E. & Jones, B. S. 1992. Conscious visual perceptual awareness vs. non-conscious visual spatial localisation examined with normal subjects using possible analogues of blindsight and neglect. *Cognitive Neuropsychology* 9:487-508.

Guzeldere, G. , Flanagan, O. , & Hardcastle, V. 2000. The nature and function of consciousness: Lessons from blindsight. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.

- Heywood, C. A. , Cowey, A. & Newcombe, F. 1991. Chromatic discrimination in a cortically colour-blind observer. *European Journal of Neuroscience* 3:802-12.
- Jackson, S. 2000. Perception, awareness and action: Insights from blindsight. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.
- Klein, S. A. 1998. Double-judgment psychophysics for research on consciousness: Application to blindsight. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Kolb, F. C. & Braun, J. 1995. Blindsight in normal observers. *Nature* 377:336-8.
- Marcel, A. J. 1998. Blindsight and shape perception: Deficit of visual consciousness or of visual function? *Brain* 121:1565-88.
- Marshall, J. C. & Halligan, P. W. 1988. Blindsight and insight in visuospatial neglect. *Nature* 336:766-67.
- Morgan, M. J. , Mason, A. J. S. & Solomon, J. A. 1997. Blindsight in normal subjects? *Nature* 385:401-2.
- Natsoulas, T. 1982. Conscious perception and the paradox of "blind-sight". In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.
- Natsoulas, T. 1997. Blindsight and consciousness. *American Journal of Psychology* 110:1-33.
- Paillard, J. , Michel, F. & Stelmach, C. E. 1983. Localization without content: A tactile analogue of "blind sight". *Archives of Neurology* 40:548-51.
- Place, U. 2000. Consciousness and the zombie within: A functional analysis of the blindsight evidence. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.
- Sahraie, A. , Weiskrantz, L. , Barbur, J. L. , Simmons, A. , & Brammer, M. 1997. Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Sciences USA* 94:9406-9411.
- Stoerig, P. & Cowey, A. 1989. Wavelength sensitivity in blindsight. *Nature* 342:916-18.
- Stoerig, P. & Cowey, A. 1989. Wavelength sensitivity in blindsight. Wavelength sensitivity in blindsight. *Brain* 115:425-44.
- Stoerig, P. & Cowey, A. 1991. Increment threshold spectral sensitivity in blindsight: Evidence for colour

opponency.

Stoerig, P. & Cowey, A. 1993. Blindsight and perceptual consciousness: Neuropsychological aspects of striate cortical function. In (B. Gulyas, D. Ottoson, & P. Roland, eds) *Functional Organization of the Human Visual Cortex*. Pergamon Press.

Stoerig, P. & Cowey, A. 1997. Blindsight in man and monkey. *Brain* 120:535-59.

Stoerig, P. 1998. Varieties of vision: From blind responses to conscious recognition. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Tapp, J. D. 1997. Blindsight in hindsight. *Consciousness and Cognition* 6:67-74.

Torjussen, T. 1978. Visual processing in cortically blind hemifields. *Neuropsychologia* 16:5-21.

Vision, G. 1998. Blindsight and philosophy. *Philosophical Psychology* 11:137-59.

Weiskrantz, L. 1986. *Blindsight: A Case Study and Implications*. Oxford University Press.

Weiskrantz, L. 1995. Blindsight: Not an island unto itself. *Current Directions in Psychological Science* 4:146-151.

Weiskrantz, L. 1995. Blindsight: Conscious vs. unconscious aspects. In (J. King & K. Pribram, eds) *Scale in Conscious Experience*. Lawrence Erlbaum.

Weiskrantz, L. 1996. Blindsight revisited. *Current Opinion in Neurobiology* 6:215-220.

Weiskrantz, L. 1997. *Consciousness Lost and Found*. Oxford University Press.

Weiskrantz, L. 1998. Consciousness and commentaries. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Weiskrantz, L. & Cowey, A. 1970. Filling in the scotoma: A study of residual vision after striate cortex lesions in monkeys. (in (E. Stellar & J. Sprague, eds) *Progress in Physiological Psychology*. Academic Press.

Weiskrantz, L. , Barbur, J. L. & Sahraie, A. 1995. Parameters affecting conscious versus unconscious visual discrimination without V1. *Proceedings of the National Academy of Sciences USA* 92:6122-26.

Weiskrantz, L. 2000. Blindsight: Implications for the conscious experience of emotion. In (R. Lane, L. Nadel, & G. Ahern, eds) *Cognitive Neuroscience of Emotion*. Oxford University Press.

Wessinger, C. M. , Fendrich, R. , Ptito, A. , & Villemure, J. 1996. Residual vision with awareness in the field contralateral to a partial or complete functional hemispherectomy. *Neuropsychologia* 34:1129-1137.

Wessinger, C. M. , Fendrich, R. , Gazzaniga, M. S. 1997. Islands of residual vision in hemianopic patients. *Journal of Cognitive Neuroscience* 9:203-21.

Zeki S. 1996. The motion vision of the blind and the modularity of consciousness. *Transactions of the Medical Society of London* 112:11-18.

Zeki, S. & ffytche, D. H. 1998. The Riddoch syndrome: Insights into the neurobiology of conscious vision. *Brain* 121:25-45.

Zihl, J. 1980. "Blindsight": Improvement of visually guided eye movements by systematic practice in patients with cerebral blindness. *Neuropsychologia* 18:71-77.

Zihl, J. & von Cramon, D. 1980. Registration of light stimuli in the cortically blind hemifield and its effect on localization. *Behavior and Brain Research* 1:287-298.

Zihl, J. & Werth, R. 1984. Contributions to the study of "blindsight", parts I & II. *Neuropsychologia* 22:1-22.

6.1d

Other Neuropsychological Disorders

Barr, W. B. 1998. Neurobehavioral disorders of awareness and their relevance to schizophrenia. In (X. Amador & A. David, eds) *Insight and Psychosis*. Oxford University Press

Bates, D. & Cartlidge, N. 1994. Disorders of consciousness. In (E. Critchley, ed) *The Neurological Boundaries of Reality*. Farrand.

Berti, A. & Rizzolatti, G. 1992. Visual processing without awareness: Evidence from unilateral neglect. *Journal of Cognitive Neuroscience* 4:345-51.

Bisiach, E. , Luzzatti, C. & Perani, D. 1979. Unilateral neglect, representational schema, and consciousness. *Brain* 102:609-18.

Bisiach, E. & Rusconi, M. L. 1990. Breakdown of perceptual awareness in unilateral neglect. *Cortex* 26:643-49.

Bisiach, E. & Geminiani, G. 1991. Anosognosia related to hemiplegia and hemianopia. In (G. Prigatano

& D. Schacter, eds) *Awareness of Deficits after Brain Injury*. Oxford University Press.

Bisiach, E. 1992. Understanding consciousness: Clues from unilateral neglect and related disorders. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Bisiach, E. 1993. Mental representation in unilateral neglect and related disorders. *Quarterly Journal of Experimental Psychology* 46A:435-61.

Bisiach, E. & Berti, A. 1995. Consciousness in dyschiria. In (M. Gazzniga, ed) *The Cognitive Neurosciences*. MIT Press.

Brady, J. P. & Lind, D. L. 1961. Experimental analysis of hysterical blindness. *Archives of General Psychiatry* 4:331-39.

De Giorgio, C. M. & Lew, M. F. 1991. Consciousness, coma, and the vegetative state: Physical basis and definitional character. *Issues in Law and Medicine* 6:361-371.

de Haan, E. H. F. , Young, A. W. , & Newcombe, F. 1987. Face recognition without awareness. *Cognitive Neuropsychology* 4:385-415.

de Renzi, E. 1986. Current issues in prosopagnosia. In (H. Ellis, M. Jeeves, F. Newcombe, & A. Young, eds) *Aspects of Face Processing*. Martinus Nijhoff.

Driver, J. & Vuilleumier, P. 2001. Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition* 79:39-88.

Engelien, A. , Huber, W. , Silbersweig, D. , Frith, C. , & Frachowiak, R. 2000. The neural correlates of 'deaf-hearing' in man. *Brain* 123:532-545.

Farah, M. J. 1990. *Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision*. MIT Press.

Farah, M. J. 1994. Perception and awareness after brain damage. *Current Opinion in Neurobiology* 4:252-55.

Farah, M. J. 1994. Visual perception and visual awareness after brain damage: A tutorial overview. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance* 15. MIT Press.

Farah, M. J. , O'Reilly, R. C. & Vecera, S. P. 1997. The neural correlates of perceptual awareness: Evidence from covert recognition in prosopagnosia. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

- Farah, M. J. & Feinberg, T. E. 1997. Consciousness of perception after brain damage. *Seminars in Neurology* 17:145-52.
- Farah, M. & Feinberg, T. 2000. Disorders of perception and awareness. In (M. Farah & T. Feinberg, eds) *Patient-based Approaches to Cognitive Neuroscience*. MIT Press.
- Farah, M. 2001. Consciousness. In (B. Rapp, ed) *The Handbook of Cognitive Neuropsychology: What Deficits Reveal about the Human Mind*. Psychology Press/Taylor & Francis.
- Feinberg, T. E. 1997. Some interesting perturbations of the self in neurology. *Seminars in Neurology* 17:129-35.
- Fredericks, J. A. M. 1969. Consciousness. In (P. Vinken & G. Bruyn, eds) *Handbook of Clinical Neurology*. North Holland.
- Frith, C. , Blakemore, S. J. , & Wolpert, D. 2000. Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews* 31:357-363.
- Galin, D. 1992. Theoretical reflections on awareness, monitoring, and self in relation on anosognosia. *Consciousness and Cognition* 1:152-62.
- Giacino J. T. 1997. Disorders of consciousness: differential diagnosis and neuropathologic features. *Seminars in Neurology* 17:105-11.
- Gibson, K. R. 1992. Toward an empirical basis for understanding consciousness and self-awareness. *Consciousness and Cognition* 1:163-68.
- Grosz, H. J. & Zimmerman, J. A. 1965. Experimental analysis of hysterical blindness: A follow-up report and new experimental data. *Archives of General Psychiatry* 13:255-60.
- Hellman, K. M. 1991. Anosognosia: Possible neuropsychological mechanisms. In (G. Prigatono & D. Schacter, eds) *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Oxford University Press.
- Humphreys, G. W. , Troscianko, T. , Riddoch, M. J. , & Boucart, M. 1992. Covert processing in different visual recognition systems. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.
- Jehkonen, M. , Ahonen, J. , Dastidar, P. , & Vilkki, J. 2000. Unawareness of deficits after right hemisphere stroke: Double-dissociations of anosognosias. *Acta Neurologica Scandinavica* 102:378-384.

Jouvet, M. 1969. Coma and other disorders of consciousness. In (P. Vinken & G. Bruyn, eds) *Handbook of Clinical Neurology*. North Holland.

Katz, J. 2000. Individual differences in the consciousness of phantom limbs. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Kihlstrom, J. F. & Tobias, B. A. 1991. Anosognosia, consciousness, and the self. In (G. Prigatono & D. Schacter, eds) *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Oxford University Press.

Knight, R. T. & Grabowecky, M. 1995. Escape from linear time: Prefrontal cortex and conscious experience. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Koehler, S. & Moscovitch, M. 1997. Unconscious visual processing in neuropsychological syndromes: A survey of the literature and evaluation of models of consciousness. In (M. Rugg, ed) *Cognitive Neuroscience*. MIT Press.

Ladavas, E. , Berti, A. , & Farne, A. 2000. Dissociation between conscious and non-conscious processing in neglect. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Lane, R. D. , Ahern, G. L. , Schwartz, G. E. & Kaszniak, A. W. 1997. Is alexithymia the emotional equivalent of blindsight? *Biological Psychiatry* 42:834-44.

Light, G. , & Braff, D. 2000. Do self-reports of perceptual anomalies reflect gating deficits in schizophrenia patients? *Biological Psychiatry* 47:463-467.

McGlynn, S. M. & Schacter, D. L. 1989. Unawareness of deficits in neuropsychological syndromes. *Journal of Clinical and Experimental Neuropsychology* 11:143-205.

Milner, A. D. & Rugg, M. (eds) 1991. *The Neuropsychology of Consciousness*. Academic Press.

Milner, A. D. 1991. Disorders of perceptual awareness: Commentary. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Newcombe, F. 1985. Neuropsychology of consciousness: A review of human clinical evidence. In (D. Oakley, ed) *Brain and Mind*. Methuen.

Porter R. J. 1991. Disorders of consciousness and associated complex behaviors. *Seminars in Neurology* 11:110-17.

- Posner J. B. 1978. Coma and other states of consciousness: the differential diagnosis of brain death. *Annals of the New York Academy of Sciences* 315:215-27.
- Prigatono, G. P. & Schacter, D. L. (eds) 1991. *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Oxford University Press.
- Ramachandran, V. S. 1995. Anosognosia in parietal lobe syndrome. *Consciousness and Cognition* 4:22-51.
- Rioch, D. M. 1954. Psychopathological and neuropathological aspects of consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.
- Schacter, D. L. , McAndrews, M. P. , and Moscovitch, M. 1986. Access to consciousness: Dissociations between implicit and explicit knowledge in neuropsychological syndromes. In (L. Weiskrantz, ed) *Thought Without Language*. Oxford University Press.
- Schacter, D. L. 1990. Toward a cognitive neuropsychology of awareness: Implicit knowledge and anosognosia. *Journal of Clinical and Experimental Neuropsychology* 12:155-78.
- Schiff, N. & Plum, F. 2000. The role of arousal and "gating" systems in the neurology of impaired consciousness. *Journal Of Clinical Neurophysiology* 17:438-452.
- Teasdale G. , Knill-Jones R. , & van der Sande J. 1978. Observer variability in assessing impaired consciousness and coma. *Journal of Neurology, Neurosurgery and Psychiatry* 41:603-10.
- Tononi, G. & Edelman, G. 2000. Schizophrenia and the mechanisms of conscious integration. *Brain Research Reviews* 31:391-400.
- Tranel, D, & Damasio, A. R. 1988. Nonconscious face recognition in patients with prosopagnosia. *Behavioral Brain Research* 30:235-49.
- Vaina, L. M. 1995. Akinetopsia, achromatopsia and blindsight: Recent studies on perception without awareness. *Synthese* 105:253-271.
- van de Kelft E. , Segnarbieux F. , Candon E. , Couchet P. , Frerebeau P. , Daures J. P. 1994. Clinical recovery of consciousness after traumatic coma. *Critical Care Medicine* 22:1108-13.
- von Cramon, D. 1978. Consciousness and disturbances of consciousness. *Journal of Neurology* 219:1-13.
- Vecera, S. P. & Gills, K. S. 1997. What is it like to be a patient with apperceptive agnosia? *Consciousness and Cognition* 6:237-66.

- Watson, R. T. , Valenstein, E. , Day, A. , & Heilman, K. M. 1994. Posterior neocortical systems subserving awareness and neglect: Neglect associated with superior temporal sulcus but not area 7 lesions. *Archives of Neurology* 51:1014-1021
- Weiskrantz, L. 1987. Neuropsychology and the nature of consciousness. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.
- Weiskrantz, L. 1988. Some contributions of neuropsychology of vision and memory to the problem of consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.
- Weiskrantz, L. 1990. Outlooks for blindsight: Explicit methodologies for implicit processes. *Proceedings of the Royal Society of London* B239:247-78.
- Weiskrantz, L. 1994. Neuropsychology and the nature of consciousness. In (H. Gutfreund & G. Toulouse, eds) *Biology and Computation: A Physicist's Choice*. World Scientific.
- Yamadori, A. 1997. Body awareness and its disorders. In (M. Ito, Y. Miyashita, & E. T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.
- Young, A. W. 1994. Covert recognition. In (M. Farah & G. Ratcliff, eds) *The Neuropsychology of High-Level Vision*. Lawrence Erlbaum.
- Young, A. W. 1994. Conscious and unconscious recognition of familiar faces. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance* 15. MIT Press.
- Young, A. W. 1995. Neuropsychology of awareness. In (A. Revonsuo & M. Kampinnen, eds) *Consciousness in Philosophy and Cognitive Neuroscience*. Lawrence Erlbaum.
- Young, A. W. 1995. Face recognition and awareness after brain injury. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.
- Young, A. W. 1996. Dissociable aspects of consciousness. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.
- Zappulla R. A. 1997. Epilepsy and consciousness. *Seminars in Neurology* 17:113-19.

6.1e

Cerebral Hemispheres and Consciousness

- Albert M. L. , Silverberg R. , Reches A. , & Berman M. 1976. Cerebral dominance for consciousness. *Archives of Neurology* 33:453-4.
- Austin, G. , Hayward, W. , & Rouhe, S. 1974. A note on the problem of conscious man and cerebral disconnection by hemispherectomy. In (M. Kinsbourne & W. Smith, eds) *Hemispheric Disconnection and Cerebral Function*. Charles C. Thomas.
- Battro, A. 2001. *Half a Brain is Enough: The Story of Nico*. Cambridge University Press.
- Baynes, K. & Gazzaniga, M. 2000. Consciousness, introspection, and the split-brain: The two minds/one body problem. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.
- Beaumont, J. 1981. Split brain studies and the duality of consciousness. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 2*. Academic Press.
- Bogen, J. E. 1968. The other side of the brain: An appositional mind. *Bulletin of the Los Angeles Neurological Society* 34:135-62.
- Bogen, J. E. 1977. Further discussion of split brains and hemispheric capabilities. *British Journal for the Philosophy of Science* 28:281-6.
- Dewitt, L. 1975. Consciousness, mind, self: The implications of the split-brain studies. *British Journal for the Philosophy of Science* 27:41-47.
- Dimond, S. J. 1978. Depletion of awareness and double-simultaneous stimulation in split-brain man. *Cortex* 14:604-607.
- Gazzaniga, M. 1977. On dividing the self: Speculations from brain research. *Excerpta Medica: Neurology* 434:233-44.
- Gazzaniga, M. S. , LeDoux, J. E. , Wilson, D. H. 1977. Language, praxis, and the right hemisphere: Clues to some mechanisms of consciousness. *Neurology* 27:1144-1147.
- Gazzaniga, M. S. 1995. Consciousness and the cerebral hemispheres. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Gazzaniga, M. & Miller, M. 2000. Testing Tulving: The split brain approach. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.
- Harrington, A. 1985. Nineteenth-century ideas on hemisphere differences and "duality of mind." *Behavioral and Brain Sciences* 8:617-660.

- Joseph, R. 1988. The right cerebral hemisphere: Emotion, music, visual-spatial skills, body-image, dreams, and awareness. *Journal of Clinical Psychology* 44:630-673.
- Kavcic, V. , Fei, R. , Hu, S. , & Doty, R. 2000. Hemispheric interaction, metacontrol, and mnemonic processing in split-brain macaques. *Behavioural Brain Research* 111:71-82.
- Kurian, G. , Santhakumari, K. 1990. Consciousness and the left cerebral hemisphere. *Journal of Indian Psychology* 8:33-36.
- LeDoux, J. E. , Wilson, D. H. & Gazzaniga, M. S. 1977. A divided mind: Observations of the conscious properties of the separated hemispheres. *Annals of Neurology* 2:417-21.
- LeDoux, J. E. , Wilson, D. H. , & Gazzaniga, M. S. 1979. Beyond commissurotomy: Clues to consciousness. In (M. Gazzaniga, ed) *Handbook of Behavioral Neurobiology*, volume 2. Plenum Press.
- LeDoux, J. E. 1986. Brain, mind, and language. In (D. Oakley, ed) *Brain and Mind*. Methuen.
- Landis, T. , Graves, R. , & Goodglass, H. 1981. Dissociated awareness of manual performance on two different visual associative tasks: A "split-brain" phenomenon in normal subjects? *Cortex* 17:435-440.
- Lishman, W. A. 1971. Emotion, consciousness, and will after brain bisection in man. *Cortex* 7:181-92.
- Mackay, D. M. 1987. Divided brains -- divided minds? In (C. Blakemore and S. Greenfield, eds) *Mindwaves*. Blackwell.
- Mark, V. 1996. Conflicting communication in a split-brain patient: Support for dual consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Marks, C. 1980. *Commissurotomy, Consciousness, and Unity of Mind*. MIT Press.
- Miller, L. 1986. Some comments on cerebral hemispheric models of consciousness. *Psychoanalytic Review* 73:129-44.
- Natsoulas, T. 1987. Consciousness and commissurotomy: 1. Spheres and Streams of consciousness. *Journal of Mind and Behavior* 8:435-468.
- Natsoulas, T. 1988. Consciousness and commissurotomy: 2. Some Pertinencies for Intact Functioning. *Journal of Mind and Behavior* 9:515-548.
- Natsoulas, T. 1991. Consciousness and commissurotomy: 3. Toward the improvement of alternative conceptions. *Journal of Mind and Behavior* 12:1-32.

- Natsoulas, T. 1992. Consciousness and commissurotomy: 4. Three hypothesized dimensions of disconnected left-hemispheric consciousness. *Journal of Mind and Behavior* 13:37-67.
- Natsoulas, T. 1991. Consciousness and commissurotomy: 5. Concerning a hypothesis of normal dual consciousness. *Journal of Mind and Behavior* 14:179-202.
- Natsoulas, T. 1991. Consciousness and commissurotomy: 6. Evidence for normal dual consciousness. *Journal of Mind and Behavior* 16:181-205.
- Preilowski B. 1979. Self-recognition as a test of consciousness in left and right hemisphere of "split-brain" patients. *Activitas Nervosa Superior* 19 (supp):343-44.
- Puccetti, R. 1977. Bilateral organization of consciousness in man. *Annals of the New York Academy of Sciences* 299:448-58.
- Puccetti, R. 1981. The case for mental duality: Evidence from split-brain data and other considerations. *Behavioral and Brain Sciences* 4:93-123.
- Quen, J. M. (ed) 1986. *Split Minds/Split Brains: Historical and Current Perspectives*. New York University Press.
- Sergent, J. 1987. A new look at the human split brain. *Brain* 110:1375-92.
- Sperry, R. W. 1968. Hemisphere disconnection and unity in conscious awareness. *American Psychologist* 23:723-733.
- Sperry, R. W. 1977. Forebrain commissurotomy and conscious awareness. *Journal of Medicine and Philosophy* 2:101-26.
- Sperry, R. W. , Zaidel, E. , Zaidel, D. 1979. Self recognition and social awareness in the disconnected minor hemisphere. *Neuropsychologia* 17:153-166.
- Sperry, R. W. 1984. Consciousness, personal identity and the divided brain. *Neuropsychologia* 22:611-73.
- Trevarthen, C. 1974. Analysis of central activities that generate and regulate consciousness in commissurotomy patients. In (S. Dimond & J. Beaumont, eds) *Hemisphere Function in the Human Brain*. Elek.
- Wessinger, C. M. , Fendrich, R. , Ptito, A. , & Villemure, J. G. 1996. Residual vision with awareness in the field contralateral to a partial or complete functional hemispherectomy. *Neuropsychologia* 34:1129-1137.

Wilkes, K. V. 1978. Consciousness and commissurotomy. *Philosophy* 53:185-99.

Zangwill, O. L. 1974. Consciousness and the cerebral hemispheres. In (S. Dimond & J. Beaumont, eds) *Heremisphere Function in the Human Brain*. Wiley.

6.1f

Neural Timing and Consciousness (Libet, etc)

Churchland, P. S. 1981. On the alleged backward referral of experience and its relevance to the mind-body problem. *Philosophy of Science* 48:165-81.

Churchland, P. S. 1981. The timing of sensations: Reply to Libet. *Philosophy of Science* 48:492-7.

Dennett, D. C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15:183-201.

Elitzur, A. 1996. Time and consciousness: The uneasy bearing of relativity on the mind-body problem. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Glynn, I. M. 1990. Consciousness and time. *Nature* 348:477-79.

Green, C. & Gillett, G. 1995. Are mental events preceded by their physical causes? *Philosophical Psychology* 8:333-340.

Honderich, T. 1984. The time of a conscious sensory experience and mind-brain theories. *Journal of Theoretical Biology* 110:115-129.

Kiefer, M. & Spitzer, M. 2000. Time course of conscious and unconscious semantic brain activation. *Neuroreport* 11:2401-2407.

Libet, B. 1978. Neuronal vs. subjective timing for a conscious sensory experience. In (P. Buser & A. Rougeul-Buser, eds) *Cerebral Correlates of Conscious Experience*. Elsevier.

Libet, B. Wright, E. W, Feinstein, B. & Pearl, D. K. 1979. Subjective referral of the timing for a cognitive sensory experience. *Brain* 102:193-224.

Libet, B. 1981. The experimental evidence for subjective referral of a sensory experience backwards in time: Reply to P. S. Churchland. *Philosophy of Science*, 48, 182-97.

Libet, B. 1981. Timing of cerebral processes relative to concomitant conscious experiences in man. In (G. Adam, I. Meszaros & E. I. Banyai, eds), *Advances in Physiological Science*. Pergamon.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8:529-66.

Libet, B. 1985. Subjective antedating of a sensory experience and mind-brain theories: Reply to Honderich. *Journal of Theoretical Biology* 114:563-70.

Libet, B. , Wright, E. W. , Feinstein, B. & Pearl, D. K. 1992. Retroactive enhancement of a skin sensation by a delayed cortical stimulus in man: Evidence for delay of a conscious sensory experience. *Consciousness and Cognition* 1:367-75.

Libet, B. 1993. The neural time factor in conscious and unconscious events. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

Libet, B. 1993. *Neurophysiology of Consciousness: Selected Papers and New Essays*. Birkhauser.

Mele, A. R. 1997. Strength of motivation and being in control - Learning from Libet. *American Philosophical Quarterly* 34:319-32.

Rossi, E. L. 1988. Paradoxes of time, consciousness, and free will: Integrating Bohm, Jung, and Libet on ethics. *Psychological Perspectives* 19:50-55.

6.1g Neural Synchrony and Binding

Cotterill, R. M. & Nielsen, C. 1991. A model for cortical 40-hertz oscillations invokes inter-area interactions. *Neuroreport* 2:289-92.

Crick, F. & Koch, C. 1990. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2:263-275.

Damasio, A. R. 1989. The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation* 1:123-32.

Damasio, A. R. 1989. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recognition and recall. *Cognition* 3:25-62.

Damasio, A. R. 1990. Synchronous activation in multiple cortical regions: A mechanism for recall. *Seminars in the Neurosciences* 2:287-96.

- Eckhorn, R. , Bauer, R. , Jordan, W. , Brosch, M. , & Reitbock, H. J. 1988. Coherent oscillations: A mechanism for feature linking in the visual cortex. *Biological Cybernetics* 60:121-30.
- Eckhorn, R. , Reitbock, H. J. , Arndt, M. , & Dicke, P. 1989. A neural network for feature linking via synchronous activity: Results from cat visual cortex and from simulations. In (R. Cotterill, ed) *Models of Brain Function*. Cambridge University Press.
- Engel, A. K. , Konig, P. & Singer, W. 1991. Direct physiologic evidence for scene segmentation by temporal coding. *Proceedings of the National Academy of Sciences USA* 88:1936-40.
- Engel, A. K. , Konig, P. Kreiter, A. K. & Schillen, T. B. 1992. Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neurosciences* 15:218-26.
- Engel, A. K. , Fries, P. , Konig, P. , Brecht, M. & Singer, W. 1999. Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition* 8:128-51.
- Engel, A. K. , Fries, P. , Konig, P. , Brecht, M. & Singer, W. 1999. Does time help to understand consciousness? *Consciousness and Cognition* 8:260-68.
- Fries, P. , Roelfsema, P. , Engel, A. , & Singer, W. 1997. Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proceedings of the National Academy of Sciences USA* 94:12699-12704.
- Gold, I. 1999. Does 40-Hz oscillation play a role in visual consciousness? *Consciousness and Cognition* 8:186-95.
- Golledge, H. D. R. , Hilgetag, C. C. , & Tovee, M. J. 1996. Information processing: A solution to the binding problem. *Current Biology* 6:1092-95.
- Gray, C. M. , Konig, P. , Engel, A. K. & Singer, W. 1992. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338:334-7.
- Gray, C. M. 1994. Synchronous oscillations in neuronal systems: Mechanisms and functions. *Journal of Computational Neuroscience* 1:11-38.
- Hardcastle, V. G. 1994. Psychology's "binding problem" and possible neurobiological solutions. *Journal of Consciousness Studies* 1:66-90.
- Hardcastle, V. G. 1996. How we get there from here: Dissolution of the binding problem. *Journal of Mind and Behavior* 17:251-66.
- Hardcastle, V. G. 1997. Consciousness and the neurobiology of perceptual binding. *Seminars in*

Neurology 17:163-70.

Konig, P. & Engel, A. K. 1995. Correlated firing in sensory-motor systems. *Current Opinion in Neurobiology* 5:511-19.

Konig, P. , Engel, A. K. & Singer, W. 1995. Relation between oscillatory activity and long-range synchronization in cat visual cortex. *Proceedings of the National Academy of Sciences USA* 92:290-94.

Konig, P. , Engel, A. K. , Roelfsema, P. R. & Singer, W. 1995. How precise is neural synchronization? *Neural Computation* 7:469-85.

Llinas, R. & Ribary, U. 1998. Temporal conjunction in thalamocortical transactions. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Nelson, J. I. 1995. Binding in the visual system. In (M. Arbib, ed) *Handbook of Brain Theory and Neural Networks*. MIT Press.

Newman, J. & Grace, A. A. 1999. Binding across time: The selective gating of frontal and hippocampal systems modulating working memory and attentional states. *Consciousness and Cognition* 8:196-212.

Prinzmetal, W. 1981. Principles of feature integration in visual perception. *Perception and Psychophysics* 30:330-40.

Revonsuo, A. 1999. Binding and the phenomenal unity of consciousness. *Consciousness and Cognition* 8:173-85.

Sauve, K. 1999. Gamma-band synchronous oscillations: Recent evidence regarding their functional significance. *Consciousness and Cognition* 8:213-24.

Schillen, T. B. & Konig, P. 1994. Binding by temporal structure in multiple feature domains of an oscillatory neural network. *Biological Cybernetics* 5:397-405.

Shastri, L. & Ajjanagadde, V. 1993. From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic binding using temporal synchrony. *Behavioral and Brain Sciences* 16:417-51.

Sillito, A. M. , Jones, H. E. , Gerstein, G. L. , & West, D. C. 1994. Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature* 369:479-82.

Singer, W. 1993. Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology* 55:349-74.

Singer, W. & Gray, C. M. 1995. Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* 18:555-86.

Singer, W. , Engel, A. K. , Kreiter, A. , Munk, M. , & Roelfsema, P. 1997. Neuronal assemblies: Necessity, signature, and detectability. *Trends in Cognitive Sciences* 1:252-60.

Steriade, M. , McCormick, D. A. , & Sejnowski, T. J. 1993. Thalamocortical oscillations in the sleeping and aroused brain. *Science* 262:679-85.

Steriade, M. 1998. Corticothalamic networks, oscillations, and plasticity. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Stryker, M. 1989. Is grandmother an oscillation? *Nature* 338:297-8.

Treisman, A. 1980. A feature integration theory of attention. *Cognitive Psychology* 12:97-136.

Treisman, A. 1996. The binding problem. *Current Opinion in Neurobiology* 6:171-8.

von der Malsburg, C. 1995. Binding in models of perception and brain function. *Current Opinion in Neurobiology* 5:520-28.

Usher, M. & Donnelly, N. 1998. Visual synchrony affects binding and segmentation in perception. *Nature* 394:179-82.

Wolfe, J. M. & Bennett, S. C. 1997. Preattentive object files: Shapeless bundles of basic features. *Vision Research* 37:25-43.

6.1h

Consciousness and Anesthesia

Aitkenhead, A. R. 1993. Conscious awareness. In (P. Sebel, B. Bonke, & E. Winograd, eds) *Memory and Awareness in Anesthesia*. Prentice-Hall.

Alkire, M. , Haier, R. , & Fallon, J. 2000. Toward a unified theory of narcosis: Brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness. *Consciousness & Cognition* 9:370-386.

Andrade, J. 1995. Learning during anesthesia: A review. *British Journal of Psychology* 86:479-506.

- Andrade, J. 1997. Investigations of hypesthesia: Using anesthetics to explore relationships between consciousness, learning, and memory. *Consciousness and Cognition* 5:562-80.
- Andrade, J. & Jones, J. G. 1997. Awareness in anesthesia. In (G. Hall & M. Morgan, eds) *Short Practice of Anesthesia*. Chapman and Hall.
- Andrade, J. 2000. NMDA receptor--mediated consciousness: A theoretical framework for understanding the effects of anesthesia on cognition? In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Bonke, B. , Fitch, W. & Millar, K. (eds) 1990. *Memory and Awareness In Anesthesia*. Swets & Zeitlinger.
- Bonke, B. , Bovill, J. G. , & Moerman, N. (eds) 1996. *Memory and Awareness in Anesthesia III*. Van Gorcum.
- Caseley-Rondi, G. , Merikle, P. M. & Bowers, K. S. 1994. Unconscious cognition in the context of general anesthesia. *Consciousness and Cognition* 3:166-95.
- Cogliolo, P. , Romano, V. , Villani, R. , & Galano, M. 1993. Effectiveness of Evans' technique for the evaluation of awareness. In (P. Sebel, B. Bonke, & E. Winograd, eds) *Memory and Awareness in Anesthesia 2*. Prentice-Hall.
- Corner, M. 1976. The nature of consciousness: some persistent conceptual difficulties and a practical suggestion. *Progress in Brain Research* 45:471-5.
- Eich, E. , Reeves, J. L. , & Katz, R. L. 1985. Anesthesia, amnesia, and the memory/awareness distinction. *Anesthesia and Analgesia* 64:1143-48.
- Evans, J. M. 1987. Patient's experiences of awareness during general anesthesia. In (M. Rosen & J. Lunn, eds) *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.
- Flohr, H. 1995. An information-processing theory of anesthesia. *Neuropsychologia* 33:1169-80.
- Flohr, H. 1998. On the mechanism of action of anesthetic agents. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Flohr, H. 2000. NMDA-receptor-mediated computational processes and phenomenal consciousness. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Franks, N. P. & Lieb, W. R. 1998. The molecular basis of general anesthesia: Current ideas. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

- Franks, N. & Lieb, W. 2000. The role of NMDA receptors in consciousness: What we learn from anesthetic mechanisms? In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Ghoneim, M. M. & Block, R. I. 1992. Learning and consciousness during general anesthesia. *Anesthesiology* 76:279-305.
- Hagan, S. , Jibu, M. & Yasue, K. 1994. Consciousness and anesthesia: A hypothesis involving biophoton emission in the microtubular cytoskeleton of the brain. In (K. Pribram, ed) *Origins: Brain and Self-organization*. Lawrence Erlbaum.
- Hameroff, S. 2001. Anesthesia: The "other side" of consciousness. *Consciousness and Cognition* 10:217-229.
- Hill, D. S. , & Hill, D. S. 1910. The loss and recovery of consciousness under anesthesia. *Psychological Bulletin* 7:77-83.
- Jansen, C. K. , Bonke, B. , Klein, J. & Bezstarosti, J. 1990. Unconscious perception during balanced anesthesia? In (B. Bonke, W. Fitch, & K. Millar, 1990) *Memory and Awareness in Anesthesia*. Swets & Zeitlinger.
- Jones, J. G. 1988. Awareness during anesthesia. *Anaesthesia Rounds*.
- Kihlstrom, J. F. & Schacter, D. L. 1990. Anesthesia, amnesia, and the cognitive unconscious. In (B. Bonke, W. Fitch, & K. Millar, 1990) *Memory and Awareness in Anesthesia*. Swets & Zeitlinger.
- Kihlstrom, J. F. & Couture, L. J. 1992. Awareness and information processing during general anesthesia. *Journal of Psychopharmacology* 6:410-17.
- Kiviniemi K. 1994. Conscious awareness and memory during general anesthesia. *Aana Journal* 62:441-9.
- Kulli, J. & Koch, C. 1991. Does anaesthesia cause loss of consciousness? *Trends in Neuroscience*, 14, 6-10.
- Levinson, B. W. 1965. States of awareness during general anaesthesia. *British Journal of Anaesthesia* 37:544-546.
- Lewis, S. A. , Jenkinson, J. & Wilson, J. 1973. An EEG investigation of awareness during anaesthesia. *British Journal of Psychology* 64:413-5.
- Merikle, P. M. , & Daneman, M. 1996. Memory for unconsciously perceived events: Evidence from anesthetized patients. *Consciousness and Cognition* 5:525-541.

- Merikle, P. M. & Daneman, M. 1996. Memory for events during anesthesia: A meta-analysis. In (B. Bonke, J. G. Bovill, & N. Moerman, eds) *Memory and Awareness in Anesthesia III*. Van Gorcum.
- Moerman, N, Bonke, B. & Oosting, J. 1993. Awareness and recall during general anesthesia: Facts and feelings. *Anesthesiology* 79:454-64.
- Mostert, J. W. 1975. States of awareness during general anesthesia. *Perspectives in Biology and Medicine* 19:68-76.
- Munglani, R. & Jones, J. G. 1992. Sleep and general anesthesia as altered states of consciousness. *Journal of Psychopharmacology* 6:399-409.
- Nikolinakos, D. 1994. General anesthesia, consciousness, and the skeptical challenge. *Journal of Philosophy* 2:88-104.
- Plourde, G. 2001. Identifying the neural correlates of consciousness: Strategies with general anesthetics. *Consciousness and Cognition* 10:241-44.
- Rosen, M. , & Lunn, J. N. (eds) 1987. *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.
- Sebel, P. S. , Bonke, B. & Winograd, E. (eds) 1993. *Memory and Awareness in Anesthesia 2*. Prentice-Hall.
- Tinnin, L. 1994. Conscious forgetting and subconscious remembering of pain. *Journal of Clinical Ethics* 5:151-52.
- Tracy, J. 1993. Awareness in the operating room: A patient's view. In (P. Sebel, B. Bonke, & E. Winograd, eds) *Memory and Awareness in Anesthesia*. Prentice-Hall.
- Utting, J. E. 1987. Awareness: Clinical aspects. In (M. Rosen & J. Lunn, eds) *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.
- White, D. C. 1987. Anesthesia: A privation of the senses: An historical introduction and some definitions. In (M. Rosen & J. Lunn, eds) *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.
- Vickers, M. D. 1987. Detecting consciousness by clinical means. In (M. Rosen & J. Lunn, eds) *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.

6.1i

Foundational Issues

- Baars, B. J. 2001. The brain basis of a "consciousness monitor":Scientific and medical significance. *Consciousness and Cognition* 10:159-164.
- Bisiach, E. 1988. The (haunted) brain and consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.
- Block, N. 1998. How to find the neural correlate of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness 1996*. MIT Press.
- Buck, R. 1993. What is this thing called subjective experience? Reflections on the neuropsychology of qualia. *Neuropsychology* 7:490-99.
- Chalmers, D. J. 1998. On the search for the neural correlate of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Chalmers, D. J. 2000. What is a neural correlate of consciousness? In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Churchland, P. S. 1988. Reduction and the neurobiological basis of consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.
- Churchland, P. S. 1994. Can neurobiology teach us anything about consciousness? *Proceedings and Addresses of the American Philosophical Association* 67:23-40.
- Cleeremans, A. & Haynes, J. 1999. Correlating consciousness: A view from empirical science. *Revue Internationale de Philosophie* 3:387-420.
- Cobb, S. 1952. On the nature and locus of mind. *Archives of Neurology and Psychiatry* 67:172-7.
- Creutzfeld, O. D. 1987. Inevitable deadlocks of the brain-mind discussion. In (B. Gulyas, ed) *The Brain-Mind Problem: Philosophical and Neurophysiological Approaches*. Leuven University Press.
- Farber, I. B. & Churchland, P. S. 1995. Consciousness and the neurosciences: Philosophical and theoretical issues. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Freeman, W. J. 1997. Three centuries of category errors in studies of the neural basis of consciousness and intentionality. *Neural Networks* 10:1175-83.
- Gillett, G. 1988. Consciousness and brain function. *Philosophical Psychology* 1:325-39.

- Gillett, G. 1995. Consciousness, thought, and neurological integrity. *Journal of Mind and Behavior* 16:215-33.
- Gloor, P. 1986. Consciousness as a neurological concept in epileptology: a critical review. *Epilepsia* 27:S14-26.
- Gordon, G. , Maxwell, G. & Savodnik, I. (eds) 1976. *Consciousness and the Brain: A Scientific and Philosophical Inquiry*. Plenum.
- Gray, J. 1998. Creeping up on the hard question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Hamanaka, T. 1997. The concept of consciousness in the history of neuropsychiatry. *History of Psychiatry* 8:361-373.
- Hardcastle, V. G. 1995. *Locating Consciousness*. John Benjamins.
- Hardcastle, V. G. 1996. Discovering the moment of consciousness? I: Bridging techniques at work, & II: An ERP analysis of priming using novel visual stimuli. *Philosophical Psychology* 9:149-96.
- Hardcastle, V. 2000. How to understand the N in NCC. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Ivanitsky A. M. 1993. Consciousness: criteria and possible mechanisms. *International Journal of Psychophysiology* 14:179-87.
- Kinsbourne, M. 1997. What qualifies a representation for a role in consciousness? In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Lahav, R. 1993. What neuropsychology tells us about consciousness. *Philosophy of Science* 60:67-85.
- Lahav, R. 1997. The conscious and the nonconscious: Philosophical implications of neuropsychology. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.
- Newton, N. 1991. Consciousness, qualia, and re-entrant signaling. *Behavior and Philosophy* 19:21.
- Niedermeyer E. 1994. Consciousness: Function and definition. *Clinical Electroencephalography* 25:86-93.
- Revonsuo, A. 1998. How to take consciousness seriously in cognitive neuroscience. *Communication and*

Cognition 30:185-205.

Revonsuo, A. 2000. Prospects for a scientific research program on consciousness. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Schiller, F. 1952. Consciousness reconsidered. *Archives of Neurology and Psychiatry* 67:199-227.

6.1j

Consciousness and Neuroscience, Misc

Arhem, P. 1996. Vertical information flow in the brain: on neuronal micro events and consciousness. *Biosystems* 38:191-98.

Bakhman, T. 2000. *Microgenetic Approach to the Conscious Mind*. John Benjamins.

Beck, H. 1976. Neuropsychological servosystems, consciousness, and the problem of embodiment. *Behavioral Science* 21:139-60.

Bernhaut, M. , Gellhorn, E. & Rasmussen, A. T. 1953. Experimental contributions to the problem of consciousness. *Journal of Neurophysiology* 16:21-35.

Boitano, J. 1996. Edelmans's biological theory of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Borrett, D. , Kelly, S. , & Kwan, H. 2000. Phenomenology, dynamical neural networks and brain function. *Philosophical Psychology* 13:213-228.

Calvin, W. 1990. *The Cerebral Symphony: Seashore Reflections on the Structure of Consciousness*. Bantam.

Cowey, A. 1997. Current awareness: Spotlight on consciousness. *Developmental Medicine and Child Neurology* 39:54-62.

Creutzfeld, O. D. 1979. Neurophysiological mechanisms and consciousness. In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.

Crick, F. & Koch, C. 1992. The problem of consciousness. *Scientific American* 267(3):152-60.

Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. Scribners.

- Damasio, A. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.
- Dehaene, S. & Naccache, L. 2001. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79:1-37.
- Delacour, J. 1995. An introduction to the biology of consciousness. *Neuropsychologia* 33:1061-1074.
- Delacour, J. 1997. Neurobiology of consciousness: An overview. *Behavioural Brain Research* 85:127-141.
- Delafresnaye, J. F. (ed) 1954. *Brain Mechanisms and Consciousness*. Blackwell.
- Desmedt J. E. & Tomberg, C. 1995. Consciousness. *Electroencephalography and Clinical Neurophysiology*, Supplement 44:227-34.
- Donald, M. 1995. The neurobiology of human consciousness: An evolutionary approach. *Neuropsychologia* 33:1087-1102.
- Donchin, E. , McCarthy, G. , Kutas, M. & Ritter, W. 1983. Event-related brain potentials in the study of consciousness. In (R. Davidson, S. Schwartz, & D. Shapiro, eds) *Consciousness and Self-Regulation*. Plenum Press.
- Donnelly G. F. 1982. Consciousness: the brain and self-regulation modalities. *Topics in Clinical Nursing* 3:13-20.
- Doty, R. W. 1975. Consciousness from neurons. *Acta Neurobiologiae Experimentalis* 35:791-804.
- Eccles, J. C. (ed) 1966. *Brain and Conscious Experience*. Springer.
- Eccles, J. C. 1974. Cerebral activity and consciousness. In (F. Ayala & T. Dobzhansky, eds) *Studies in the Philosophy of Biology*. University of California Press.
- Eccles, J. C. 1987. The effect of silent thinking on the cerebral cortex. In (B. Gulyas, ed) *The Brain-Mind Problem: Philosophical and Neurophysiological Approaches*. Leuven University Press.
- Edelman, G. M. 1989. *The Remembered Present: A Biological Theory of Consciousness*. Basic Books.
- Edelman, G. M. 1992. *Bright Air, Brilliant Fire: On the Matter of the Mind*. Penguin.
- Fessard, A. E. 1952. Mechanisms of nervous integration and conscious experience. In (J. Delafresnaye,

ed) *Brain Mechanisms and Consciousness*. Blackwell.

Frith, C. D. 1992. Consciousness, information processing, and the brain. *Journal of Psychopharmacology* 6:436-40.

Frith, C. D. 1996. The role of the prefrontal cortex in self-consciousness: the case of auditory hallucinations. *Philosophical Transactions of the Royal Society of London* B351:1505-12.

Gastaut, H. 1954. The brain stem and cerebral electrogenesis in relation to consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.

Gazzaniga, M. 1988. Brain modularity: Toward a philosophy of conscious experience. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Globus, G. , Maxwell, G. & Savodnik, I. (eds) 1975. *Consciousness and the Brain*. Plenum Press.

Greenfield, S. 1995. *Journey to the Centers of the Mind*. W. H. Freeman.

Jasper, H. & Shagass, C. 1941. Conscious time judgments related to conditioned time intervals and voluntary control of the alpha rhythm. *Journal of Experimental Psychology* 28:503-508.

Kety, S. S. 1952. Consciousness and the metabolism of the brain. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Third Conference*. Josiah Macy Foundation.

Kinsbourne, M. 1995. Models of consciousness: Serial or parallel in the brain? In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Kokoszka, A. 1993. Information metabolism as a model of consciousness. *International Journal of Neuroscience* 68:165-77.

Lin, S. , Tsai, Y. , & Liou, C. 1993. Conscious mental tasks and their EEG signals. *Medical and Biological Engineering and Computing* 31:421-26.

Luria, A. 1976. The human brain and conscious activity. In Schwartz & Shapiro 1978.

Mitterauer B. 1998. An interdisciplinary approach towards a theory of consciousness. *Biosystems* 45:99-121.

Pare, D. & Llinas, R. 1995. Conscious and pre-conscious processes as seen from the standpoint of sleep-waking cycle neurophysiology. *Neuropsychologia* 33:1155-1168.

- Penfield, W. 1975. *The Mystery of the Mind*. Princeton University Press.
- Petty, P. G. 1998. Consciousness: A neurosurgical perspective. *Journal of Consciousness Studies* 5:86-96.
- Picton, T. W. & Stuss, D. T. 1994. Neurobiology of conscious experience. *Current Opinion in Neurobiology* 4:256-65.
- Pribram, K. H. 1990. Brain and consciousness: A wealth of data. In (E. John, ed), *Machinery of the Mind: Data, Theory, and Speculations about Higher Brain Function*. Boston: Birkhauser.
- Ramachandran, V. S. & Hirstein, W. 1998. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies* 4:429-57.
- Rose, S. 1973. *The Conscious Brain*. Paragon House.
- Rudell, A. P. & Hua, J. 1996. The recognition potential and conscious awareness. *Electroencephalography and Clinical Neurophysiology* 98:309-18.
- Salazar, A. , Grafman J. , Vance S. , Dillon J. D. , & Ludlow, C. 1986. Consciousness and amnesia after penetrating head injury: neurology and anatomy. *Neurology* 36:178-87.
- Singer, W. 2000. Phenomenal awareness and consciousness from a neurobiological perspective. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Simonov, P. V. 1994. Consciousness and the brain. *Neuroscience and Behavioral Physiology* 24:234-38.
- Travis F. T. & Orme-Johnson D. W. 1989. Field model of consciousness: EEG coherence changes as indicators of field effects. *International Journal of Neuroscience* 49:203-11.
- Trevarthen, C. 1979. The tasks of consciousness: How could the brain do them? In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.
- Turner B. H. & Knapp M. E. 1995. Consciousness: A neurobiological approach. *Integrative Physiological and Behavioral Science* 30:151-6.
- Walter, W. G. 1954. Theoretical properties of diffuse projection systems in relation to behaviour and consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.
- Webb, A. C. 1970. Consciousness and the cerebral cortex. *British Journal of Anaesthesia* 55:209-19.
- Zeman, A. Z. J. , Grayling, A. C. & Cowey, A. 1997. Contemporary theories of consciousness. *Journal of*

Neurology, Neurosurgery, and Psychiatry 62:549-552.

6.2

Consciousness and Psychology

6.2a

Cognitive Models of Consciousness

Allport, D. A. 1979. Conscious and unconscious cognition: A computational metaphor for the mechanism of attention and integration. In (L. Nilsson, ed) *Perspectives on Memory Research*.

Baars, B. J. 1983. Conscious contents provide the nervous system with coherent, global information. In (R. Davidson, G. Schwartz, & D. Shapiro, eds) *Consciousness and Self-Regulation*. Plenum.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baars, B. J. 1993. How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity? In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174).

Baars, B. J. , Fehling, M. R. , LaPolla, M. , & McGovern, K. 1997. Consciousness creates access: Conscious goal images recruit unconscious action routines, but goal competition serves to "liberate" such routines, causing predictable slips. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Baars, B. J. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press.

Baars, B. J. 1997. In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies* 4:292-309.

Baars, B. J. 1998. Metaphors of consciousness and attention in the brain. *Trends in Neurosciences* 21:58-62.

Baars, B. J. & McGovern, K. 1996. Cognitive views of consciousness: What are the facts? How can we explain them? In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Bechtel, W. 1995. Consciousness: Perspectives from symbolic and connectionist AI. *Neuropsychologia*.

- Brown, R. A. 1997. Consciousness in a self-learning, memory-controlled, compound machine. *Neural Networks* 10:1333-85.
- Browne, C. , Evans, R. , Sales, N. , & Aleksander, I. L. 1997. Consciousness and neural cognizers: A review of some recent approaches. *Neural Networks* 10:1303-1316.
- Burks, A. W. 1986. An architectural theory of functional consciousness. In (N. Rescher, ed) *Current Issues in Teleology*. University Press of America.
- Cabanac, M. 1996. On the origin of consciousness, a postulate, and its corollary. *Neuroscience and Biobehavioral Reviews* 20:33-40.
- Cam, P. 1989. Notes toward a faculty theory of cognitive consciousness. In (P. Slezak, ed) *Computers, Brains and Minds*. Kluwer.
- Carr, T. H. 1979. Consciousness in models of human information processing: Primary memory, executive control, and input regulation. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 1*. Academic Press.
- Cotterill, R. M. J. 1996. Prediction and internal feedback in conscious perception. *Journal of Consciousness Studies* 3:245-66.
- Cotterill, R. M. J. 1997. On the mechanism of consciousness. *Journal of Consciousness Studies* 4:231-48.
- Cotterill, R. M. J. 1997. Navigation, consciousness and the body/mind "problem". *Psyke and Logos*, 18:337-341.
- Farrell, B. A. 1970. The design of a conscious device. *Mind* 79:321-46.
- Gregory R. L. 1984. Is consciousness sensational inferences? *Perception* 13:641-6.
- Hardcastle, V. G. 1995. A critique of information processing theories of consciousness. *Minds and Machines* 5:89-107.
- Harnad, S. 1982. Consciousness: An afterthought. *Cognition and Brain Theory* 5:29-47.
- Harth, E. 1993. *The Creative Loop: How the Brain Makes a Mind*. Addison Wesley.
- Harth, E. 1995. The sketchpad model: A theory of consciousness, perception, and imagery. *Consciousness and Cognition* 4:346-68.

- Harth, E. 1996. Self-referent mechanisms as the neuronal basis of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Jackendoff, R. 1987. *Consciousness and the Computational Mind*. MIT Press.
- John, E. R. 1976. A model of consciousness. In (G. Schwartz & D. Shapiro, eds) *Consciousness and Self-Regulation*. Plenum Press.
- Johnson, M. K. & Reeder, J. A. 1997. Consciousness as meta-processing. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Johnson-Laird, P. 1983. A computational analysis of consciousness. *Cognition and Brain Theory* 6:499-508. Also in (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.
- Kawato, M. 1997. Bidirectional theory approach to consciousness. In (M. Ito, Y. Miyashita, & E. T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.
- Lauro-Grotto, R. , Reich, S. & Virasoro, M. A. 1997. The computational role of conscious processing in a model of semantic memory. In (M. Ito, Y. Miyashita, & E. T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.
- Lloyd, D. 1995. Consciousness: A connectionist manifesto. *Minds and Machines* 5:161-85.
- Lloyd, D. 1996. Consciousness, connectionism, and cognitive neuroscience: A meeting of the minds. *Philosophical Psychology* 9:61-78.
- Mathis, D. W. & Moxer, M. 1995. On the computational utility of consciousness. In (G. Tesauro, D. Touretzky, & T. Leen, eds) *Advances in Neural Information Processing Systems* 7. MIT Press.
- Michie, D. 1994. Consciousness as an engineering issue (Parts 1 and 2). *Journal of Consciousness Studies* 1:192-95, 2:52-66.
- Norretranders, T. 1991. *The User Illusion: Cutting Consciousness Down to Size*. Viking Penguin.
- O'Brien, G. & Opie, J. 1998. A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences* 22:127-48.
- Oatley, K. 1981. Representing ourselves: Mental schemata, computational metaphors, and the nature of consciousness. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 2*. Academic Press.

- Parsons, T. 1953. Consciousness and symbolic processes. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Fourth Conference*. Josiah Macy Foundation.
- Phaf, R. H. & Wolters, G. 1997. A constructivist and connectionist view on conscious and nonconscious processes. *Philosophical Psychology* 10:287-307.
- Restian, A. 1981. Informational analysis of consciousness. *International Journal of Neuroscience* 13:229-37.
- Revonsuo, A. 1993. Cognitive models of consciousness. In (M. Kamppinen, ed) *Consciousness, Cognitive Schemata, and Relativism*. Kluwer.
- Roberts, H. 1968. Consciousness in animals and automata. *Psychological Reports* 22:1226-28.
- Rolls, E. T. 1997. Consciousness in neural networks? *Neural Networks* 10:1227-1303.
- Schacter, D. L. 1989. On the relation between memory and consciousness: Dissociable interactions and conscious experience. In (H. Roediger & F. Craik, eds) *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*.
- Schneider, W. & Pimm-Smith, M. 1997. Consciousness as a message-aware control mechanism to modulate cognitive processing. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Shallice, T. 1972. Dual functions of consciousness. *Psychological Review* 79:383-93.
- Shallice, T. 1978. The dominant action system: An information-processing approach to consciousness. In (K. S. Pope & J. L. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.
- Shallice, T. 1988. Information-processing models of consciousness: possibilities and problems. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.
- Sommerhoff G. & MacDorman K. 1994. An account of consciousness in physical and functional terms: A target for research in the neurosciences. *Integrative Physiological and Behavioral Science* 29:151-81.
- Sommerhoff, G. 1996. Consciousness as an internal integrating system. *Journal of Consciousness Studies* 3:139-57.
- Strehler, B. L. 1989. Monitors: key mechanisms and roles in the development and aging of the consciousness and self. *Mechanisms of Ageing and Development* 47:85-132.

Sun, R. 1997. Learning, action, and consciousness: A hybrid approach toward modeling consciousness. *Neural Networks* 10:1317-33.

Sviderskaya, N. E. 1991. Consciousness and information selection. *Neuroscience and Behavioral Physiology* 21:526-31.

Taylor, J. G. 1996. Modeling what it is like to be. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Taylor, J. G. 1996. A competition for consciousness? *Neurocomputing* 11:271-96.

Taylor, J. G. 1997. Neural networks for consciousness. *Neural Networks* 10:1207-27.

Taylor, J. G. & Mueller-Gaertner, H. 1997. Non-invasive analysis of awareness. *Neural Networks* 10:1185-1194.

Werbos, P. 1997. Optimization: A foundation for understanding consciousness. In (D. Levine & W. Elsberry, eds) *Optimality in Biological and Artificial Networks?*. Lawrence Erlbaum.

6.2b

Unconscious Perception

Adams, J. K. 1957. Laboratory studies of behavior without awareness. *Psychological Bulletin* 54:383-405.

Balay, J. & Shevrin, H. 1988. The subliminal psychodynamic activation method: A critical review. *American Psychologist* 43:161-74.

Bar, M. 2000. Conscious and nonconscious processing of visual object identity. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Bargh, J. A. 1992. Does subliminality matter to social psychology? Awareness of the stimulus versus awareness of its influence. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.

Bhalla, M. & Proffitt, D. 2000. Geographical slant perception: Dissociation and coordination between explicit awareness and visually guided actions. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Bornstein, R. F. 1989. Exposure and affect: Overview and meta-analysis of research 1968-1987.

Psychological Bulletin 106:265-89.

Bornstein, R. F. 1992. Subliminal mere exposure effects. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.

Bornstein, R. F. & Pittman, T. S. 1992. *Perception without Awareness: Cognitive, Clinical, and Social Perspectives*. Guilford.

Bowers, K. S. 1982. On being unconsciously influenced and informed. In (K. Bowers & D. Meichenbaum, eds) *The Unconscious Reconsidered*. Wiley.

Cheesman, J. & Merikle, P. M. 1984. Priming with and without awareness. *Perception and Psychophysics* 36:387-95.

Cheesman, J. & Merikle, P. M. 1986. Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology* 40:343-67.

Damian, M. 2001. Congruity effects evoked by subliminally presented primes: Automaticity rather than semantic processing. *Journal of Experimental Psychology: Human Perception & Performance* 27:154-165.

Debnar, J. A. & Jacoby, L. L. 1994. Unconscious perception: Attention, awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20:304-17.

Dixon, N. F. 1971. *Subliminal Perception: The Nature of a Controversy*. McGraw-Hill.

Dixon, N. F. & Henley, S. H. A. 1980. Without awareness. In (M. Jeeves, ed) *Psychology Survey* 3. Allen and Unwin.

Doyle, J. R. 1990. Detectionless processing with semantic activation? A footnote to Greenwald, Klinger, and Liu (1989). *Memory and Cognition* 18:428-9.

Eagle, M. 1959. The effects of subliminal stimuli of aggressive content upon conscious cognition. *Journal of Personality* 27:578-600.

Erdelyi, M. H. 1970. Recovery of unavailable perceptual input. *Cognitive Psychology* 1:99-113.

Eriksen, C. W. 1956. An experimental analysis of subception. *American Journal of Psychology* 69:625-34.

Eriksen, C. W. 1956. Subception: Fact or artifact? *Psychological Review* 63:74-80.

- Eriksen, C. W. 1960. Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review* 67:279-300.
- Farah, M. J. , Monheit, M. A. & Wallace, M. A. 1991. Unconscious perception of "extinguished" visual stimuli: Reassessing the evidence. *Neuropsychologia* 29:949-58.
- Fowler, C. A. , Woldford, G. , Slade, R. & Tassinary, L. 1981. Lexical access with and without awareness. *Journal of Experimental Psychology: General* 110:341-62.
- Fuhrer, M. J. & Eriksen, C. W. 1960. The unconscious perception of the meaning of verbal stimuli. *Journal of Abnormal and Social Psychology* 61:432-9.
- Goldiamond, I. 1958. Indicators of perception: 1. Subliminal perception, subception, unconscious perception: An analysis in terms of psychophysical indicator methodology. *Psychological Bulletin* 55:373-411.
- Greenwald, A. G. , Klinger, M. R. , & Liu, T. J. 1989. Unconscious processing of dichoptically masked words. *Memory and Cognition* 17:35-47.
- Greenwald, A. G. , Spangenberg, E. , Pratkanis, A. R. , & Eskenazi, J. 1991. Double blind tests of subliminal self-help audiotapes. *Psychological Science* 2:119-22.
- Greenwald, A. G. , Klinger, M. R. & Schuh, E. S. 1995. Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General* 124:22-42.
- Greenwald, A. G. & Draine, S. 1997. Do subliminal stimuli enter the mind unnoticed?: Tests with a new method. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Hardaway, R. A. 1990. Subliminally activated symbiotic fantasies: Facts and artifacts. *Psychological Bulletin* 107:177-95.
- Henley, S. H. 1984. Unconscious perception revisited: A comment on Merikle (1992). *Bulletin of the Psychonomic Society* 22:121-4.
- Holender, D. 1986. Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences* 9:1-23.
- Kemp-Wheeler, S. M. & Hill, A. B. 1988. Semantic priming without awareness: Some methodological considerations and implications. *Quarterly Journal of Experimental Psychology* 40:671-92.

- Khurana, B. 2000. Face representation without conscious processing. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Kihlstrom, J. F. 1996. Perception without awareness of what is perceived, learning without awareness of what is learned. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.
- Kihlstrom, J. F. , Barnhardt, T. M. & Tataryn, D. J. 1992. Implicit perception. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.
- Klauer, K. & Greenwald, A. 2000. Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception & Performance* 26:1506-1508.
- Kostandov, E. A. 1994. Subsensory reactions and the problem of unconscious perception. *Sensory Systems* 7:149-53.
- Krosnick, J. A. , Betz, A. L. , Jussim, L. J. & Lynn, A. R. 1992. Subliminal conditioning of attitudes. *Personality and Social Psychology Bulletin* 18:152-62.
- Kunst-Wilson, W. R. & Zajonc, R. B. 1980. Affective discrimination of stimuli that cannot be recognized. *Science* 207:557-58.
- Kunzendorf, R. G. 1985. Subconscious percepts as "unmonitored" percepts: An empirical study. *Imagination, Cognition and Personality* 4:365-73.
- Lazarus, R. S. & McCleary R. A. 1983. Autonomic discrimination without awareness: A study of subception. *Psychological Review* 58:113-22.
- Lewicki, P. , Hill, T. & Czyewska, M. 1992. Nonconscious acquisition of information. *American Psychologist* 47:792-801.
- Lewis, J. L. 1970. Semantic processing of unattended messages using dichotic listening.
- MacLeod, C. 1998. Implicit perception: Perceptual processing without awareness. In (K. Kirsner, G. Speelman, eds) *Implicit and Explicit Mental Processes*. Lawrence Erlbaum.
- Marcel, A. J. 1983. Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology* 15:197-237.
- Marcel, A. J. 1983. Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology* 15:238-300.

- Merikle, P. M. 1982. Unconscious perception revisited. *Perception and Psychophysics* 31:298-301.
- Merikle, P. M. & Reingold, E. M. 1990. Recognition and lexical decision without detection: Unconscious perception? *Journal of Experimental Psychology: Human Perception and Performance* 16:574-83.
- Merikle, P. M. 1992. Perception without awareness: Critical issues. *American Psychologist* 47:792-5.
- Merikle, P. M & Reingold, E. M. 1992. Measuring unconscious processes. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.
- Merikle, P. M. , Joordens, S. & Stolz, J. A. 1995. Measuring the relative magnitude of unconscious influences. *Consciousness and Cognition* 4:422-39.
- Merikle, P. M. & Daneman, M. 1997. Psychological investigations of unconscious perception. *Journal of Consciousness Studies*.
- Merikle, P. & Daneman, M. 2000. Conscious vs. unconscious perception. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.
- Merikle, P. M. , D. Smilek, & J. D. Eastwood 2001. Perception without awareness: Perspectives from cognitive psychology. *Cognition* 79:115-34.
- Miller, J. 2000. Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception & Performance* 26:1461-1477.
- Moore, T. E. 1992. Subliminal perception: Facts and fallacies. *Skeptical Inquirer* 16:273-81.
- Neuberg, S. L. 1988. Behavioral implications of information presented outside of conscious awareness: The effect of subliminal presentation of trait information on behavior in the Prisoner's Dilemma game. *Social Cognition* 6:207-30.
- Nolan, K. A. & Caramazza, A. 1982. Unconscious perception of meaning: A failure to replicate. *Bulletin of the Psychonomic Society* 20:23-26.
- Peirce, C. S. & Jastrow, J. 1884. On small differences in sensation. *Memoirs of the National Academy of Sciences* 3:75-83.
- Pisella, L. & Rosetti, Y. 2000. Interaction between conscious identification and non-conscious sensory-motor processing: Temporal constraints. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

- Poppel, E. , Held, R. & Frost, D. 1973. Residual function after brain wounds involving the central visual pathways in man. *Nature* 243:295-96.
- Pratkanis, A. R. & Greenwald, A. G. 1988. Recent perspectives on unconscious processing: Still no marketing applications. *Psychology and Marketing* 5:337-53.
- Purcell, D. G. , Stewart, A. L. & Stanovich, K. K. 1983. Another look at semantic priming without awareness. *Perception and Psychophysics* 34:65-71.
- Reingold, E. M. & Merikle, P. M. 1988. Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics* 44:563-575.
- Reingold, E. & Merikle, P. 1991. Theory and measurement in the study of unconscious processes. *Mind and Language* 5:9-28.
- Schmidt, T. 2000. Visual perception without awareness: priming responses by color. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Shevrin, H. 1992. Unconscious perception, memory, and consciousness: Cognitive and dynamic perspectives. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.
- Silverman, L. H. & Weinberger, J. 1985. Mommy and I are one: Implications for psychotherapy. *American Psychologist* 40:1296-1308.
- Theus, K. T. 1994. Subliminal advertising and the psychology of processing unconscious stimuli: A review of research. *Psychology and Marketing* 11:271-290.
- Young, A. & Ellis, H. 2000. Overt and covert face recognition. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.
- Weinberger, J. & Hardaway, R. 1990. Separating science from myth in subliminal psychodynamic activation. *Clinical Psychological Review* 10:727-56.

6.2c

Unconscious Processes

- Abrams, R. & Greenwald, A. 2000. Parts outweigh the whole (word) in unconscious analysis of meaning. *Psychological Science* 11:118-124.
- Bridgeman, B. 1992. Conscious vs unconscious processes: The case of vision. *Theory and Psychology*

2:73-88.

Buchner, A. 1997. Consciousness, intention, and the process dissociation procedure. *Sprache and Kognition* 16:176-182.

Cowan, N. , Stadler, M. A. 1996. Estimating unconscious processes: Implications of a general class of models. *Journal of Experimental Psychology: General* 125:195-200.

Dixon, N. F. 1981. *Preconscious Processing*. Wiley.

Erdelyi, M. H. 1974. A new look at the New Look: Perceptual defense and vigilance. *Psychological Review* 81:1-25.

Erdelyi, M. H. 1992. Psychodynamics and the unconscious. *American Psychologist* 47:784-87.

Field, A. 2000. I like it, but I'm not sure why: Can evaluative conditioning occur without conscious awareness? *Consciousness & Cognition* 9:13-36.

Ford, T. & Thompson, E. 2000. Preconscious and postconscious processes underlying construct accessibility effects: An extended search model. *Personality & Social Psychology Review* 4:317-336.

Gaito, J. 1964. Stages of perception, unconscious processes, and information extraction. *Journal of General Psychology* 70:183-197.

Greenwald, A. G. 1992. New Look 3: Unconscious cognition reclaimed. *American Psychologist* 47:766-79.

Hilgard, E. R. 1958. *Unconscious Processes and Man's Rationality*. University of Illinois Press.

Hoffman, R. 1997. What neural network studies suggest regarding the boundary between conscious and unconscious mental processes. In (D. Stein, ed) *Cognitive Science and the Unconscious*. American Psychiatric Press

Hommel, B. 2000. Intentional control of automatic stimulus-response translation. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Imanaka, K. & Abernethy, B. 2000. Distance-location interference in movement reproduction: An interaction between conscious and unconscious processing? In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Jacoby, L. L. , Toth, J. P. , Lindsay, D. .S. , Debner, J. A. 1992. Lectures for a layperson: Methods for

- revealing unconscious processes. In (R. Bornstein & B. Pittman, eds) *Perception without Awareness: Cognitive, Clinical, and Social Perspectives*. Guilford Press.
- Kihlstrom, J. F. 1984. Conscious, subconscious, unconscious: A cognitive perspective. In (K. S. Bowers & D. Meichenbaum, eds) *The Unconscious Reconsidered*. Wiley.
- Kihlstrom, J. F. 1987. The cognitive unconscious. *Science* 237:1445-1452.
- Kihlstrom, J. F. 1990. The psychological unconscious. In (L. Pervin, ed) *Handbook of Personality: Theory and Research*. Guilford Press.
- Kihlstrom, J. F. , Barnhardt, T. M. & Tattryn, D. J. 1992. The psychological unconscious: Found, lost, and regained. *American Psychologist* 47:788-91.
- Kihlstrom, J. F. 1995. The rediscovery of the unconscious mind. In (H. Morowitz & J. Singer, eds) *The Mind, the Brain, and Complex Adaptive Systems*. Addison-Wesley.
- Kihlstrom, J. F. 1996. Unconscious processes in social interaction. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Klinger, M. , Burton, P. , & Pitts, G. 2000. Mechanisms of unconscious priming: Response competition, not spreading activation. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 26:441-455.
- Lewicki, P. 1986. *Nonconscious Social Information Processing*. Academic Press.
- Lewicki, P. & Hill, T. 1987. Unconscious processes as explanations of behavior in cognitive, personality, and social psychology. *Personality and Social Psychology Bulletin* 13:355-362.
- Loftus, E. F. & Klinger, M. R. 1992. Is the unconscious smart or dumb? *American Psychologist* 47:761-65.
- Miller, J. G. 1951. Unconscious processes and perception. In (R. Blake & G. Ramsey, eds) *Perception*.
- Miller, J. G. 1952. The experimental study of unconscious processes. In (M. Reymert, ed) *Feelings and Emotions*.
- Perner, J. & Clements, W. 2000. From an implicit to an explicit "theory of mind". In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

- Peterfreund, E. & Schwartz, J. T. 1971. Information processing and the nature of conscious and unconscious processes. *Psychological Issues* 7:219-29.
- Posner, M. I. 1991. Recent experimental studies of conscious and unconscious processes. In (M. Posner, B. Dwivedi, & I. Singh, eds) *Contemporary Approaches to Cognitive Psychology*. Rishi Publications.
- Radil, T. , Radilova, J. , Bozkov, V. , & Bohdanecky, Z. 1981. Unconscious and conscious processes during visual perception. *Acta Neurobiologiae Experimentalis* 41:565-572.
- Revonsuo, A. , Johanson, M. , Wedlund, J. , & Chaplin, J. 2000. The zombies among us: Consciousness and automatic behaviour. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.
- Schacter, D. L. 1992. Implicit knowledge: New perspectives on unconscious processes. *Proceedings of the National Academy of Sciences USA* 89:11113-17.
- Schwartz, M. 1981. Criteria for physiological substrates of unconscious processes. *American Psychologist* 36:434-435.
- Shevrin, H. & Dickman, S. 1980. The psychological unconscious: A necessary assumption for all psychological theory? *American Psychologist* 35:421-34.
- Shevrin, H. , Smith, W. H. , Fitzler, D. E. 1971. Average evoked response and verbal correlates of unconscious mental processes. *Psychophysiology* 8:149-62.
- Shevrin, H. & Fritzler, D. E. 1968. Visual evoked response correlates of unconscious mental processes. *Science* 161:295-298.
- Spitz, H. H. 1993. The role of the unconscious in thinking and problem solving. *Educational Psychology* 13:229-244.
- Spitz, H. H. 1995. Calendar calculating idiots savants and the smart unconscious. *New Ideas in Psychology* 13:167-182.
- Stein, D. J. (ed) 1997. *Cognitive Science and the Unconscious*. American Psychiatric Press.
- Suzuki, K. & Yamadori, A. 2000. Intact verbal description of letters with diminished awareness of their forms. *Journal of Neurology, Neurosurgery & Psychiatry* 68:782-786.
- Underwood, G. & Bright, J. E. H. 1996. Cognition with and without awareness. In (G. Underwood, ed) *Implicit Cognition*. Oxford University Press.

Wachtel, P. L. 1987. "The unconscious" and unconscious processes. *Canadian Psychology* 28:107-108.

Yu, J. & Bellezza, F. 2000. Process dissociation as source monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 26:1518-1533.

6.2d

Visual Consciousness [see also 6.1b, 6.1c, 6.2b, 6.2e, 6.2j]

Bachmann, T. 1997. Visibility of brief images: The dual-process approach. *Consciousness and Cognition* 6:491-518.

Bachmann, T. 1998. Fast dynamics of visibility of brief visual images: The perceptual-retouch viewpoint. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Baxt, N. 1982. On the time necessary for a visual impression to come into consciousness. *Psychological Research* 44:1-12.

Blackmore, S. J. , Brelstaff, G. , Nelson, K. & Troscianko, T. 1995. Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. *Perception* 24:1075-81.

Bridgeman, B. , Hendry, D. & Stark, L. 1975. Failure to detect displacements of the visual world during saccadic eye movements. *Vision Research* 15:719-22.

Bridgeman, B. 2000. Interactions between vision for perception and vision for behavior. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Cogan, A. I. 1995. Vision comes to mind. *Perception* 24:811-26.

Di Lollo, V. , Enns, J. , & Rensink, R. 2000. Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal Of Experimental Psychology-General* 129:481-507.

Dixon, M. , Smilek, D. , Cudahy, C. , & Merikle, P. 2000. Five plus two equals yellow: Mental arithmetic in people with synaesthesia is not coloured by visual experience. *Nature* 406:365.

Durgin, F. H. 1995. On the filling in of the visual blind spot: Some rules of thumb. *Perception* 24:827-40.

Fernandez-Duque, D. & Thornton, I. 2000. Change detection without awareness: Do explicit reports underestimate the representation of change in the visual system? *Visual Cognition* 7:323-344.

- Grimes, J. 1996. On the failure to detect changes in scenes across saccades. In (K. Akins, ed) *Perception*. Oxford University Press.
- Irwin, D. E. 1991. Information integration across saccadic eye movements. *Cognitive Psychology* 23:420-56.
- Iwasaki, S. 1993. Spatial attention and two modes of visual consciousness. *Cognition* 49:211-233.
- Lachter, J. , Durgin, F. , & Washington, T. 2000. Disappearing percepts: Evidence for retention failure in metacontrast masking. *Visual Cognition* 7:269-279.
- Mack, A. & Rock, I. 1998. *Inattentional Blindness*. MIT Press.
- McConkie, G. W. & Zola, D. 1979. Is visual information integrated across successive fixations in reading? *Perception and Psychophysics* 25:221-24.
- Natsoulas, T. 1993. An introduction to reflective seeing. *Journal of Mind and Behavior* 14:235-56.
- Natsoulas, T. 1994. An introduction to reflective seeing. *Journal of Mind and Behavior* 15:351-74.
- Nijhawan, R. & Khurana, B. 2000. Conscious registration of continuous and discrete visual events. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Noe, A. , Pessoa, L. , & Thompson, E. 2000. Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition* 7: 93-106.
- O'Regan, J. K. 1992. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology* 46:461-88.
- O'Regan, J. , Deubel, H. , Clark, J. , & Rensink, R. 2000. Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition* 7:191-211.
- Palmer, S. 1999. *Vision Science: Photons to Phenomenology*. MIT Press.
- Pashler, H. 1988. Familiarity and visual change detection. *Perception and Psychophysics* 41:191-201.
- Pessoa, L. , Thompson, E. & Noe, A. 1998. Finding out about filling in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*.
- Phillips, W. A. 1974. On the distinction between sensory storage and visual short-term memory. *Perception and Psychophysics* 16:283-90.

- Ramachandran, V. S. 1992. Filling in gaps in perception: Part I. *Current Directions in Psychological Science* 1:199-205.
- Rensink, R. A. , O'Regan, J. K. , & Clark, J. J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8:368-373.
- Rensink, R. , O'Regan, J. , & Clark, J. 2000. On the failure to detect changes in scenes across brief interruptions. *Visual Cognition Special* 7:127-145.
- Scott-Brown, K. , Baker, M. , & Orbach, H. 2000. Comparison blindness. *Visual Cognition* 7:253-267.
- Shapiro, K. 2000. Change blindness: Theory or paradigm? *Visual Cognition* 7:83-91.
- Shore, D. , & Klein, R. 2000. The effects of scene inversion on change blindness. *Journal of General Psychology* 127:27-43.
- Simons D. J. & Levin, D. T. 1997. Change blindness. *Trends in Cognitive Science* 1:241-82.
- Simons, D. 2000. Current approaches to change blindness. *Visual Cognition* 7:1-15.
- Sperling, G. 1960. The information available in visual presentations. *Psychological Monographs* 74:1-29.
- Wallis, G. & Buelthoff, H. 2000. What's scene and not seen: Influences of movement and task upon what we see. *Visual Cognition* 7:175-190.

6.2e Consciousness and Attention [see also 6.2d]

- Arvidson, P. S. 1996. Toward a phenomenology of attention. *Human Studies* 19:71-84.
- Baars, B. J. 1997. Some essential differences between consciousness and attention, perception, and working memory. *Consciousness and Cognition* 6:363-371.
- Baars, B. J. 1999. Attention vs. consciousness in the visual brain: Differences in conception, phenomenology, behavior, neuroanatomy, and physiology. *Journal of General Psychology* 126:224-33.
- Baddeley, A. D. & Weiskrantz, L. (eds) 1993. *Attention: Selection, Awareness, and Control*. Oxford University Press.
- Bridgeman, B. 1986. Relations between the physiology of attention and the physiology of consciousness. *Psychological Research* 48:259-266.

- Cobb, C. 1955. Awareness, attention, and physiology of the brain stem. In (P. Hoch & J. Zubin, eds) *Experimental Psychopathology*.
- Coslett, H. B. 1997. Consciousness and attention. *Seminars in Neurology* 17:137-44.
- Cowan, N. & Wood, N. L. 1997. Constraints on awareness, attention, processing, and memory: Some recent investigations with ignored speech. *Consciousness and Cognition* 6:182-203.
- Csikszentmihalyi, M. 1978. Attention and the holistic approach to behavior. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.
- Hardcastle, V. G. 1997. Attention versus consciousness: A distinction with a difference. Manuscript.
- He, S. , Cavanagh, P. & Intriligator, J. 1996. Attentional resolution and the locus of visual awareness. *Nature* 383:334-37.
- Hochberg, J. 1970. Attention, organization, and consciousness. In (D. Mostofsky, ed) *Attention: Contemporary Theory and Analysis*. Appleton-Century-Crofts.
- Iwasaki, S. 1993. Spatial attention and two modes of visual consciousness. *Cognition* 49:211-233.
- LaBerge, D. 1997. Attention, awareness, and the triangular circuit. *Consciousness and Cognition* 9:149-81.
- LaBerge, D. , Auclair, L. , & Sieroff, E. 2000. Preparatory attention: Experiment and theory. *Consciousness & Cognition* 9:396-434.
- Loper, A. B. & Hallahan, D. P. 1982. Meta-attention: The development of awareness of the attentional process. *Journal of General Psychology* 106:27-33.
- Mack, A. & Rock, I. 1998. *Inattentional Blindness*. MIT Press.
- Merikle, P. M. & Joordens, S. 1997. Parallels between perception without attention and perception without awareness. *Consciousness and Cognition* 6:219-36.
- McCormick, P. A. 1997. Orienting attention without awareness. *Journal of Experimental Psychology: Human Perception & Performance* 23:168-180.
- Newman, J. B. , Baars, B. J. , & Cho, S. 1997. A neural global workspace model for conscious attention. *Neural Networks* 10:1195-1206.

- Newman, J. B. 1995. Thalamic contributions to attention and consciousness. *Consciousness and Cognition* 4:172-93.
- Newsome, W. T. 1996. Visual attention: spotlights, highlights and visual awareness. *Current Biology* 6:357-60.
- Posner, M. I. 1994. Attention: The mechanisms of consciousness. *Proceedings of the National Academy of Sciences USA* 91:7398-7403.
- Posner, M. I. & Rothbart, M. K. 1992. Attentional mechanisms and conscious experience. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.
- Prinzmetal, W. Amiri, H. , Allen, K. & Edwards, T. 1997. The phenomenology of attention, part 1: Color, location, orientation, and "clarity". *Journal of Experimental Psychology: Human Perception and Performance*.
- Prinzmetal, W. , Nwachuku, I. , Bodanski, L. , & Blumenfeld, L. 1997. The phenomenology of attention, part 2: Brightness and contrast. *Consciousness and Cognition* 6:372-412.
- Rensink, R. A. , O'Regan, J. K. , & Clark, J. J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8:368-373.
- Rensink, R. 2000. Visual search for change: A probe into the nature of attentional processing. *Visual Cognition* 7:345-376.
- Scheier, M. F. Matthews, K. A. & Carver, C. S. 1983. Focus of attention and awareness of bodily states. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.
- Scholl, B. 2000. Attenuated change blindness for exogenously attended items in a flicker paradigm. *Visual Cognition* 7:377-396.
- Shiffrin, R. M. 1997. Attention, automatism, and consciousness. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Smilek, D. , Eastwood, J. , & Merikle, P. 2000. Does unattended information facilitate change detection? *Journal of Experimental Psychology: Human Perception & Performance* 26:480-487.
- Umiltà, C. & Moscovitch, M. 1994. *Attention and Performance 15: Conscious and Nonconscious Information Processing*. MIT Press.
- Underwood, G. 1977. Attention, awareness, and hemispheric differences in word recognition.

Neuropsychologia 15:61-67.

Underwood, G. 1983. Selective attention and selective awareness of conscious processes. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

6.2f Consciousness and Memory

Alkire, M. T. , Haier, R. J. , Fallon, J. H. , & Barker, S. J. 1996. PET imaging of conscious and unconscious verbal memory. *Journal of Consciousness Studies* 3:448-62.

Allik, J. 2000. Available and accessible information in memory and vision. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Andreasen, N. 2000. Is schizophrenia a disorder of memory or consciousness? In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Baddeley, A. 1993. Working memory and conscious awareness. In (A. Collins, S. Gathercole, M. Conway, & P. Morris, eds) *Theories of Memory*. Lawrence Erlbaum.

Balint, E. 1987. Memory and consciousness. *International Journal of Psychoanalysis* 68:475-483.

Barba, G. 2000. Memory, consciousness, and temporality: What is retrieved and who exactly is controlling the retrieval? In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Brainerd, C. J. , Stein, L. M. , & Reyna, V. F. 1998. On the development of conscious and unconscious memory. *Developmental Psychology* 34:342-357.

Brewer, W. F. 1992. Phenomenal experience in laboratory and autobiographical memory. In (M. Conway, D. Rubin, H. Spinnler, & W. Wagenaar, eds) *Theoretical Perspectives on Autobiographical Memory*. Kluwer.

Brewer, W. F. 1996. What is recollective memory? In (D. Rubin, ed) *Remembering our Past: Studies in Autobiographical Memory*. Cambridge University Press.

Buchner, A. , Erdfelder, E. , Vaterrodt-Plunnecke, B. 1995. Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General* 124:137-60.

Cavanaugh, J. C. 1989. The importance of awareness in memory aging. In (L. Poon, D. Rubin, & B. Wilson, eds) *Everyday Cognition in Adulthood and Late Life*. Cambridge University Press.

- Clark, R. E. & Squire, L. R. 1998. Classical conditioning and brain systems: The role of awareness. *Science* 280:77-81.
- Cloitre, M. 1997. Conscious and unconscious memory: A model of functional amnesia. In (D. Stein, ed) *Cognitive Science and the Unconscious*. American Psychiatric Press.
- Conway, M. A. & Dewhurst, S. A. 1995. The self and recollective experience. *Applied Cognitive Psychology* 9:1-19.
- Dalla Barba, G. 2000. Memory, consciousness, and the brain. *Brain & Cognition* 42:20-22.
- Duezel, E. 2000. What brain activity tells us about conscious awareness of memory retrieval. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.
- Duzel E. , Yonelinas A. P. , Mangun G. R. , Heinze H. J. , & Tulving E. 1997. Event related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences of the United States of America* 94:5973-8.
- Eich, E. 1984. Memory for unattended events: Remembering with and without awarene
- Erdelyi, M. 1984. The recovery of unconscious (inaccessible) memories: Laboratory studies of hypermnesia. In (G. Bower, ed) *The Psychology of Learning and Motivation*. Academic Press.
- Gardiner, J. M. 1988. Functional aspects of recollective experience. *Memory and Cognition* 16:309-13.
- Gardiner, J. M. & Parkin, A. J. 1990. Attention and recollective experience in recognition memory. *Memory and Cognition* 18:579-583.
- Gardiner, J. M. 1993. Recognition memory and awareness: An experiential approach. *European Journal of Cognitive Psychology* 5:337-46.
- Gardiner, J. M. 1996. On consciousness in relation to memory and learning. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.
- Gardiner, J. M, Ramponi, C. & Richardson-Klavehn, A. 1998. Experiences of remembering, knowing, and guessing. *Consciousness and Cognition* 7:1-26.
- Gardiner, J. & Richardson-Klavehn, A. 2000. Remembering and knowing. In (E. Tulving & Craik, F. , eds) *The Oxford Handbook of Memory*. Oxford University Press.

- Gennaro, R. J. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333-47.
- Gregg, V. H. & Gardiner, J. M. 1994. Recognition memory and awareness: A large effect of study-test modalities on "know" responses following a highly perceptual orienting task. *European Journal of Cognitive Psychology* 6:137-47.
- Hamann, S. B. & Squire, L. R. 1997. Intact perceptual memory in the absence of conscious memory. *Behavioral Neuroscience* 111:850-54.
- Hirshman, E. & Master, S. 1997. Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory and Cognition* 25:345-351.
- Hirst, W. 1989. On consciousness, recall, recognition, and the architecture of memory. In (S. Lewandowsky, J. Dunn, & K. Kirsner, eds) *Implicit Memory: Theoretical Issues*. Lawrence Erlbaum.
- Jacoby, L. L. & Witherspoon, D. 1982. Remembering without awareness. *Canadian Journal of Psychology* 36:300-324.
- Jacoby, L. L. & Kelley, C. M. 1987. Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin* 13:314-36.
- Jacoby, L. L. 1991. A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* 30:513-41.
- Jacoby, L. L. & Kelley, C. M. 1991. Unconscious influences of memory: Dissociations and automaticity. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.
- Jacoby, L. L. , Toth, J. P. & Yonelinas, A. P. 1993. Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General* 122:139-54.
- Jacoby, L. L. , Toth, J. P. , Yonelinas, A. P. & Debnar, J. A. 1994. The relation between conscious and unconscious influences: Independence or redundancy? *Journal of Experimental Psychology: General*.
- Jacoby, L. L. , Yonelinas, A. P. , & Jennings, J. M. 1997. The relation between conscious and unconscious (automatic) influences: A declaration of independence. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Johnson, M. K. , Foley, M. A. , Suengas, A. G. & Raye, C. L. 1988. Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General* 117:371-76.

- Joordens, S. & Merikle, P. M. 1993. Independence or redundancy? Two models of conscious and unconscious influences. *Journal of Experimental Psychology: General* 122:462-67.
- Kelley, C. M. & Jacoby, L. L. 1990. The construction of subjective experience: Memory attributions. *Mind and Language* 5:49-68.
- Kelley, C. M. & Lindsay, D. S. 1996. Conscious and unconscious forms of memory. In (E. Bjork & R. Bjork, eds) *Memory: Handbook of Perception and Cognition*. Academic Press.
- Kihlstrom, J. F. 1995. Memory and consciousness: An appreciation of Claparede and "Recognition et Moitte". *Consciousness and Cognition* 4:379-86.
- Kuhlmann, F. 1906. On the analysis of the memory consciousness: A study in the mental imagery and memory of meaningless visual forms. *Psychological Review* 13:316-48.
- Lampinen, J. M. , Neuschatz, J. S. , & Payne, D. G. 1998. Memory illusions and consciousness: Examining the phenomenology of true and false memories. *Current Psychology: Developmental, Learning, Personality, Social* 16:181-224.
- Larsen, S. F. 1998. What is it like to remember? On phenomenal qualities of memory. In (C. Thompson, J. Read, D. Bruce, D. Payne, & M. Toglia, eds) *Autobiographical and Eyewitness Memory: Theoretical and Applied Perspectives*. Lawrence Erlbaum.
- Light, L. L. , Singh, A. , & Capps, J. L. 1986. Dissociation of memory and awareness in young and older adults. *Journal of Clinical & Experimental Neuropsychology* 8:62-74.
- Lockhart, R. S. 1989. Consciousness and the function of remembered episodes. In (H. Roediger & F. Craik, eds) *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*.
- Mandler, G. 1989. Memory: Conscious and unconscious. In (P. Solomon, G. Goethals, C. Kelley, & R. Stephens, eds) *Memory: Interdisciplinary Approaches*. Springer-Verlag.
- Moscovitch, M. & Umiltà, C. 1991. Conscious and nonconscious aspects of memory: A neuropsychological framework of modules and central systems. In (R. Lister & H. Weingartner, eds) *Perspectives on Cognitive Neuroscience*. Oxford University Press
- Moscovitch, M. 1992. A neuropsychological model of memory and consciousness. In (L. Squire & N. Butters, eds) *Neuropsychology of Memory*. Guilford Press.
- Moscovitch, M. , Goshen-Gottstein, Y. & Vriezen, E. 1994. Memory without conscious recollection: A tutorial review from a neuropsychological perspective. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance 15*. MIT Press.

- Moscovitch, M. 1995. Recovered consciousness: A hypothesis concerning modularity and episodic memory. *Journal of Clinical and Experimental Neuropsychology*, 17:276-90.
- Moscovitch, M. 1995. Models of consciousness and memory. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Moscovitch, M. 2000. Theories of memory and consciousness. In (E. Tulving & F. Craik, eds) *The Oxford Handbook of Memory*. Oxford University Press.
- Natsoulas, T. 1986. Consciousness and memory. *Journal of Mind and Behavior* 7:463-501.
- Oakhill, J. & Kyle, F. 2000. The relation between phonological awareness and working memory. *Journal of Experimental Child Psychology* 75:152-164.
- Paller, K. A. , Kutas, M. & McIsaac, H. K. 1995. Monitoring conscious recollection via the electrical activity of the brain. *Psychological Science* 6:107-11.
- Paller, K. 2000. Neural measures of conscious and unconscious memory. *Behavioural Neurology* 12:127-141.
- Penfield, W. 1969. Consciousness, memory, and man's conditioned reflexes. In (H. Hyden, ed) *On the Biology of Learning*. Harcourt, Brace, and World.
- Rajaram, S. & Roediger, H. L. 1997. Remembering and knowing as states of consciousness during retrieval. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Rajaram, S. 1998. The effects of conceptual salience and perceptual distinctiveness on conscious recollection. *Psychonomic Bulletin and Review* 5:71-78.
- Reder, L. M. (ed) 1996. *Implicit Memory and Metacognition*. Lawrence Erlbaum.
- Reingold, E. M. & Toth, P. 1996. Memory dissociations versus task dissociations: A controversy in progress. In (G. Underwood, ed) *Implicit Cognition*. Oxford University Press.
- Richardson-Klavehn, A. & Gardner, J. M. 1996. Cross-modality priming in stem completion reflects conscious memory, but not voluntary memory. *Psychonomic Bulletin and Review* 3:238-44.
- Richardson-Klavehn, A. , Gardiner, J. M. & Java, R. I. 1996. Memory: Task dissociations, process dissociations and dissociations of consciousness. In (G. Underwood, ed) *Implicit Cognition*. Oxford University Press.

- Roberts, H. M. 1971. Conscious experiences are a memory process. *Psychological Reports* 29:591-94.
- Roediger, H. L. & Craik, F. I. M. (eds). *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*. Lawrence Erlbaum.
- Ronnberg, J. & Archer, T. 1992. Purposive behaviour in cognition and perception: Considerations of awareness in memory. *Scandinavian Journal of Psychology* 33:86-91.
- Rugg, M. D. 1995. Memory and consciousness: A selective review of issues and data. *Neuropsychologia* 33:1131-1141.
- Schacter, D. L. 1987. Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:501-18.
- Schacter, D. L. 1989. On the relation between memory and consciousness: Dissociable interactions and conscious experience. In (H. Roediger & F. Craik, eds) *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*.
- Schacter, D. L. , Bowers, J. & Booker, J. 1989. Intention, awareness, and implicit memory: The retrieval intentionality criterion. In (S. Lewandowsky, J. Dunn, & K. Kirsner, eds) *Implicit Memory: Theoretical Issues*. Lawrence Erlbaum.
- Schacter, D. L. 1992. Consciousness and awareness in memory and amnesia: Critical issues. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.
- Schacter, D. L. 1995. Implicit memory: A new frontier for cognitive neuroscience. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Schacter, D. L. 1998. Memory and awareness. *Science* 280:59-60.
- Stolz, J. & Merikle, P. 2000. Conscious and unconscious influences of memory: Temporal dynamics. *Memory* 8:333-343.
- Toth, J. P. , Lindsay, D. S, & Jacoby, L. L. 1992. Awareness, automaticity, and memory dissociations. In (L. Squire & N. Butters, eds) *Neuropsychology of Memory*. Guilford Press.
- Toth, J. P. & Reingold, E. M. 1996. Beyond perception: Conceptual contributions to unconscious influences of memory. In (G. Underwood, ed) *Implicit Cognition*. Oxford University Press.
- Tulving, E. 1985. Memory and consciousness. *Canadian Psychology* 26:1-12.

- Tulving, E. 1987. Multiple memory systems and consciousness. *Human Neurobiology* 6:67-80.
- Tulving, E. 1993. Varieties of consciousness and levels of awareness in memory. In (A. Baddeley & L. Weiskrantz, eds) *Attention: Selection, Awareness, and Control*. Oxford University Press.
- Underwood, G. 1979. Memory systems and conscious processes. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 1*. Academic Press.
- Verfaellie, M. & Keane, M. M. 1997. The neural basis of aware and unaware forms of memory. *Seminars in Neurology* 17:153-61.
- Wheeler, M. A. Stuss, D. T. & Tulving, E. 1997. Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin* 121:331-54.
- Wippich, W. 1992. Implicit and explicit memory without awareness. *Psychological Research* 54:212-24.

6.2g

Consciousness and Learning

- Berry, D. C. 1994. Implicit learning: Twenty-five years on. A tutorial. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance 15*. MIT Press.
- Berry, D. C. 1997. *How Implicit is Implicit Learning?*. Oxford University Press.
- Berry, D. C. & Dienes, Z. (eds) 1993. *Implicit Learning: Theoretical and Empirical Issues*. Lawrence Erlbaum Associates.
- Brody, N. 1989. Unconscious learning of rules: Comment on Reber's analysis of implicit learning. *Journal of Experimental Psychology: General* 118:236-238.
- Carlson, R. A. & Dulany, D. E. 1985. Conscious attention and abstraction in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11:45-58.
- Cleeremans, A. 1993. Mechanisms of implicit learning: Connectionist models of sequence processing.
- Conway, M. A. , Collins, A. F. , Anderson, S. J. , & Cohen, G. 1998. Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General*.
- Dienes, Z. & Berry, D. 1997. Implicit learning: Below the subjective threshold. *Psychonomic Bulletin &*

Review 4:3-23.

Dulany, D. E. 1968. Awareness, rules, and propositional control: A confrontation with S-R behavior theory. In (T. Dixon & D. Horton, eds) *Verbal Behavior and General Behavior Theory*. Prentice-Hall.

Dulany, D. E. , Carlson, R. A. & Dewey, G. I. 1984. A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General* 113:541-555.

Dulany, D. E. , Carlson, R. A. , Dewey, G. I. 1985. On consciousness in syntactic learning and judgment: A reply to Reber, Allen, and Regan. *Journal of Experimental Psychology: General* 114:25-32.

Furedy, J. , Damke, B. , & Boucsein, W. 2000. Revisiting the learning-without-awareness question in human Pavlovian autonomic conditioning: Focus on extinction in a dichotic listening paradigm. *Integrative Physiological & Behavioral Science* 35:17-34.

Gardiner, J. M. 1996. On consciousness in relation to memory and learning. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Kihlstrom, J. F. 1996. Perception without awareness of what is perceived, learning without awareness of what is learned. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Manns, J. , Clark, R. , & Squire, L. 2000. Awareness predicts the magnitude of single-cue trace eyeblink conditioning. *Hippocampus* 10:181-186.

Marton, F. & Booth, S. A. 1997. *Learning and Awareness*. Lawrence Erlbaum.

Neal, A. & Hesketh, B. 1997. Episodic knowledge and implicit learning. *Psychonomic Bulletin and Review* 4:24-37.

Neal, A. & Hesketh, B. 1997. Future directions for implicit learning: Toward a clarification of issues associated with knowledge representation and consciousness. *Psychonomic Bulletin and Review* 4:73-78.

Perruchet, P. , Gallego, J. & Savy, I. 1990. A critical reappraisal of the evidence for unconscious abstraction of deterministic rules in complex experimental situations. *Cognitive Psychology* 22:493-516.

Perruchet, P. & Pacteau, C. 1990. Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge. *Journal of Experimental Psychology: General* 119:264-75.

Perruchet, P. , Vinter, A. , & Gallego, J. 1997. Implicit learning shapes new conscious percepts and representations. *Psychonomic Bulletin & Review* 4:43-48.

Reber, A. S. 1967. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal*

Behavior 6:855-863.

Reber, A. S. 1989. Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* 118:219-35.

Reber, A. S. 1993. *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. Oxford University Press.

Reber, A. S. 1997. How to differentiate implicit and explicit modes of acquisition. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Reber, A. S. , Allen, R. , & Regan, S. 1985. Syntactical learning and judgment, still unconscious and still abstract: Comment on Dulany, Carlson, and Dewey. *Journal of Experimental Psychology: General* 114:17-24.

Shanks, D. R. , Green, R. E. A. , & Kolodny, J. A. 1994. A critical examination of the evidence for unconscious (implicit) learning. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance* 15. MIT Press.

Shanks, D. R. & St. John, M. F. 1994. Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* 17:367-447.

Sno, H. 2000. Deja vu and jamais vu. In (G. Berrios & J. Hodges, eds) *Memory Disorders in Psychiatric Practice*. Cambridge University Press.

Stadler, M. A. & Roediger, H. L. 1998. The question of awareness in research on implicit learning. In (M. Stadler & P. Frensch, eds) *Handbook of Implicit Learning*. Sage Publications

Stadler, M. A. , & Frensch, P. A. 1998. *Handbook of Implicit Learning*. Sage Publications.

Whittlesea, B. W. A. & Dorken, M. D. 1997. Implicit learning: Indirect, not unconscious. *Psychonomic Bulletin & Review* 4:63-67.

Wong, P. S. , Bernat, E., Bunce, S., & Shevrin, H. 1997. Brain indices of nonconscious associative learning. *Consciousness and Cognition* 6:519-544.

6.2h

Consciousness and Metacognition

Brown, R. & McNeill, D. 1966. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and*

Verbal Behavior 5:325-37.

Brown, S. 2000. Tip-of-the-tongue phenomena: An introductory phenomenological analysis. *Consciousness & Cognition* 9:516-537.

Hart, J. T. 1965. Memory and the feeling-of-knowing experience. *Journal of Educational Psychology* 56:208-16.

Johnson, M. K. 1988. Reality monitoring: An experimental phenomenological approach. *Journal of Experimental Psychology: General* 117:390-94.

Johnson, M. K. 1991. Reflection, reality monitoring, and the self. In (R. Kunzendorf, ed) *Mental Imagery*. Plenum Press.

Johnson, M. K. & Reeder, J. A. 1997. Consciousness as meta-processing. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Kahan, T. L. & LaBerge, S. 1994. Lucid dreaming as metacognition: Implications for cognitive science. *Consciousness & Cognition* 3:246-64.

Mangan, B. 2000. What feeling is the "feeling of knowing?" *Consciousness & Cognition* 9:538-544.

Metcalf, J. & Shimamura, P. 1994. *Metacognition: Knowing about Knowing*. MIT Press.

Nelson, T. O. 1996. Consciousness and metacognition. *American Psychologist* 51:102-16.

Nelson, T. O. 1992. *Metacognition: Core Readings*. Allyn and Bacon.

Nisbett, R. & Wilson, T. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review* 84:231-59.

Otani, H. & Hodge, M. 1991. mechanisms of feelings of knowing: The role of elaloration and familiarity. *Psychological Record* 41:523-35.

Reder, L. M. (ed) 1996. *Implicit Memory and Metacognition*. Lawrence Erlbaum.

Reder, L. M. & Schunn, C. D. 1996. Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In (L. Reder, ed) *Implicit Memory and Metacognition*. Lawrence Erlbaum.

Ricciardelli, L. A. 1993. Two components of metalinguistic awareness: Control of linguistic processing

and analysis of linguistic knowledge. *Applied Psycholinguistics* 14:349-367.

Rosenthal, D. M. 1998. Consciousness and metacognition. In (D. Sperber, ed) *Metarepresentation*. Oxford University Press.

Wegner, D. M. 1989. *White Bears and Other Unwanted Thoughts: Suppression, Obsession, and the Psychology of Mental Control*. Penguin.

Wegner, D. M. 1997. Why the mind wanders. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

White, P. 1980. Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review* 87:105-12.

White, P. 1983. Knowing our own minds: Conscious awareness and verbal reports. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Wilson, T. D. 1997. The psychology of metapsychology. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

6.2i

Consciousness and Control

Baars, B. J. 1987. What is conscious in the control of action? A modern ideomotor theory of voluntary action. In (D. Gorfein & R. Hoffman, eds) *Learning and Memory: The Ebbinghaus Centennial Symposium*. Lawrence Erlbaum.

Baars, B. J. 1992. *Experimental Slips and Human Error: Exploring the Architecture of Volition*. Plenum Press.

Baars, B. J. 1993. Why volition is a foundation issue for psychology. *Consciousness and Cognition* 2:281-309.

Bargh, J. A. 1996. Automaticity in action: The unconscious as repository of chronic goals and motives. In (P. Gollwitzer & J. Bargh, eds) *The Psychology of Action: Linking Cognition and Motivation to Behavior*. Guilford.

Bargh, J. A. 1996. Automaticity in social psychology. In (E. Higgins & A. Kruglanski, eds) *Social Psychology: Handbook of Basic Principles*. Guilford.

- Bargh, J. A. 1994. The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In (R. Wyer & T. Srull, eds) *Handbook of Social Cognition*. Lawrence Erlbaum.
- Bayles, G. H. & Cleary, P. J. 1986. The role of awareness in the control of frontalis muscle activity. *Biological Psychology* 22:23-35.
- Carr, T. H. , McCauley, C. , Sperber, R. D. , & Parmelee, C. M. 1982. Words, pictures, and priming: On semantic activation, conscious identification, and the automaticity of information processing. *Journal of Experimental Psychology: Human Perception & Performance* 8:757-777.
- Daprati, E. , Franck, N. , Georgieff, N. ,; Proust, J. 1997. Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition* 65:71-86.
- Delabarre, E. B. 1911. Volition and motor consciousness: Theory. *Psychological Bulletin* 8:378-82.
- Delabarre, E. B. 1912. Volition and motor consciousness: Theory. *Psychological Bulletin* 9:409-13.
- Delabarre, E. B. 1913. Volition and motor consciousness. *Psychological Bulletin* 10:441-44.
- Dewan, E. M. 1976. Consciousness as an emergent causal agent in the context of control system theory. In (G. Globus, G. Maxwell, & I. Savodnik, eds) *Consciousness and the Brain*. Plenum Press.
- Gott, P. S. , Hughes, E. C. & Whipple, K. 1984. Voluntary control of two lateralized conscious states: Validation of electrical and behavioral studies. *Neuropsychologia* 22:65-72.
- Gordon, A. M. & Rosenbaum, D. A. 1984. Conscious and subconscious arm movements: Application of signal detection theory to motor control. *Bulletin of the Psychonomic Society* 22:214-216.
- Gray, J. A. 1998. Abnormal contents of consciousness: The transition from automatic to controlled processing. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.
- Hilgard, E. R. 1979. Consciousness and control: Lessons from hypnosis. *Australian Journal of Clinical & Experimental Hypnosis* 7:103-15.
- Horowitz, M. J. & Stinson, C. H. 1995. Consciousness and processes of control. *Journal of Psychotherapy Practice and Research* 4:123-139.
- Jacoby, L. L. 1991. A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* 30:513-41.
- Jacoby, L. L. , Ste-Marie, D. & Toth, J. P. 1993. Redefining automaticity: Unconscious influences,

awareness, and control. In (A. Baddeley & L. Weiskrantz, eds) *Attention: Selection, Awareness, and Control*. Oxford University Press.

Kamiya, J. 1968. Conscious control of brain waves. *Psychology Today* 1:56-60.

Kimble, G. A. & Perlmutter, L. C. 1970. The problem of volition. *Psychological Review* 77:361-84.

Langer, E. J. 1992. Matters of mind: Mindfulness/mindlessness in perspective. *Consciousness and Cognition* 1:289-305.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8:529-566.

Oswald, M. & Gadenne, V. 2000. Are controlled processes conscious? In (W. Perrig & A. Grob, eds) *Control of Human Behavior, Mental Processes, and Consciousness: Essays in Honor of the 60th Birthday of August Flammer*. Lawrence Erlbaum Associates.

Plotkin, W. B. 1976. On the self-regulation of the occipital alpha rhythm: Control strategies, states of consciousness, and the role of physiological feedback. *Journal of Experimental Psychology: General* 105:66-99.

Plotkin, William B. 1981. A rapprochement of the operant-conditioning and awareness views of biofeedback training: The role of discrimination in voluntary control. *Journal of Experimental Psychology: General* 110:415-428.

Posner, M. & Snyder, C. R. R. 1975. Attention and cognitive control. In (R. Solso, ed) *Information Processing and Cognition: The Loyola Symposium*. Lawrence Erlbaum.

Raichle, M. E. 1997. Automaticity: From reflective to reflexive information processing. In (M. Ito, Y. Miyashita, & E. T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.

Schneider, W. E. & Shiffrin, R. M. 1977. Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* 84:1-66.

Shiffrin, R. M. & Schneider, W. E. 1977. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review* 84:128-90.

Tzelgov, J. 1997. Specifying the relations between automaticity and consciousness: A theoretical note. *Consciousness and Cognition* 6:441-51.

Tzelgov, J. , Porat, Z. & Henik, A. 1997. Automaticity and consciousness: Is perceiving the word necessary for reading it? *American Journal of Psychology* 110:429-48.

Tzelgov, J. 1997. Automatic but conscious: That is how we act most of the time. In (R. Wyer, ed) *The Automaticity of Everyday Life*. Lawrence Erlbaum.

Uleman, J. S. 1987. Consciousness and control: The case of spontaneous trait inferences. *Personality and Social Psychology Bulletin* 13:337-54.

Umiltà, C. 1988. The control operations of consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

White, W. A. 1920. Extending the field of conscious control. *Mental Hygiene* 4:857-66.

Zelazo, P. D. & Frye, D. 1997. Cognitive complexity and control: A theory of the development of deliberate reasoning and intentional action. In (M. Stamenov, ed) *Language Structure, Discourse, and the Access to Consciousness*. John Benjamins.

6.2j

Consciousness and Imagery

Ahsen, A. 1991. Imagery and consciousness: Putting together poetic, mythic and social realities. *Journal of Mental Imagery* 15:63-97.

Ahsen, A. 1991. A second report on AA-VVIQ: Role of vivid and unvivid images in consciousness research. *Journal of Mental Imagery* 15:1-31.

Ahsen, A. 1993. Imagery paradigm: Imaginative consciousness in the experimental and clinical setting. *Journal of Mental Imagery*.

Arnheim, R. 1994. Consciousness: An island of images. *Journal of Theoretical and Philosophical Psychology* 14:121-27.

Bichowsky, F. R. 1926. The mechanism of consciousness: Images. *American Journal of Psychology* 37:557-564.

Frick, R. W. 1987. A dissociation of conscious visual imagery and visual short-term memory. *Neuropsychologia* 25:707-12.

Hampson, P. J. & Morris, P. E. 1990. Imagery, consciousness, and cognitive control: The BOSS model reviewed. In (P. Hampson, D. Marks, & J. Richardson, Eds) *Imagery: Current Developments*. Routledge.

- Hebb, D. O. 1968. Concerning imagery. *Psychological Review* 75:466-77.
- Horne, P. V. 1993. The nature of imagery. *Consciousness and Cognition* 2:58-82.
- Hubbard, T. L. 1996. The importance of a consideration of qualia to imagery and cognition. *Consciousness and Cognition* 3:327-58.
- Ishai, A. & Sagi, D. 1998. Visual imagery and visual perception: The role of memory and conscious awareness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Krellenstein, M. F. 1995. Unsolvable problems, visual imagery, and explanatory satisfaction. *Journal of Mind and Behavior* 16:235-54.
- Kunzendorf, R. G. 1990. The causal efficacy of consciousness in general, imagery in particular: A materialist perspective. In (R. Kunzendorf, ed) *Mental Imagery*. Plenum Press.
- Kunzendorf, R. G. , Justice, M. , & Capone, D. 1997. Conscious images as "centrally excited sensations": A developmental study of imaginal influences on the ERG. *Journal of Mental Imagery* 21:155-66.
- Lehmann, D. , Henggler, B. , Koukkan, M. & Michel, M. 1993. Source localization of brain electric field frequency bands during conscious, spontaneous visual imagery and abstract thought. *Cognitive Brain Research* 1:203-20.
- Mandler, G. 1984. Consciousness, imagery, and emotion -- with special reference to autonomic imagery. *Journal of Mental Imagery* 8:87-94.
- Marks, D. F. 1977. Imagery and consciousness: A theoretical review from an individual differences perspective. *Journal of Mental Imagery* 1:275-90.
- Marks, D. F. 1983. Imagery and consciousness: A theoretical review. In (A. Sheikh, ed) *Imagery: Current Theory, Research, and Application*. Wiley.
- Marks, D. F. 1990. On the relationship between imagery, body, and mind. In (P. Hampson, D. Marks, & J. Richardson, eds) *Imagery: Current Developments*. Routledge.
- Mavromatis, A. 1987. On shared states of consciousness and objective imagery. *Journal of Mental Imagery* 11:125-30.
- Morris, P. E. & Hampson, P. J. 1983. *Imagery and Consciousness*. Academic Press.
- Newton, N. 1982. Experience and imagery. *Southern Journal of Philosophy* 21:475-87.

Richardson, A. 2000. Individual differences in visual imagination imagery. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Sheehan, P. W. & Lewis, S. E. 1974. Subjects' reports of confusion in consciousness and the arousal of imagery. *Perceptual and Motor Skills* 38:731-34.

6.2k

Consciousness and Emotion

Cioffi, D. 1991. Sensory awareness versus sensory impression: Affect and attention interact to produce somatic meaning. *Cognition and Emotion* 5:275-94.

Damasio, A. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.

DeLancey, C. 1996. Emotion and the function of consciousness. *Journal of Consciousness Studies* 3:492-99.

Dimberg, U. , Thunberg, M. , & Elmehed, K. 2000. Unconscious facial reactions to emotional facial expressions. *Psychological Science* 11:86-89.

Ellis, R. D. 1995. *Questioning Consciousness: The Interplay of Imagery, Cognition, and Emotion in the Human Brain*. John Benjamins.

Ellis, R. 2000. *The Caldron of Consciousness: Motivation, Affect and Self-organization*. John Benjamins.

Forgas, J. & Ciarrochi, J. 2000. Affect infusion and affect control: The interactive role of conscious and unconscious processing strategies in mood management. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Gray, J. A. 1999. Cognition, emotion, conscious experience and the brain. In (T. Dalgleish & M. J. Powers, eds) *Handbook of Cognition and Emotion*. Wiley.

Heilman, K. 2000. Emotional experience: A neurological model. In (R. Lane, L. Nadel, & G. Ahern, J. Allen, & A. Kaszniak, eds) *Cognitive Neuroscience of Emotion*. Oxford University Press.

Katz, J. M. 1983. Altered states of consciousness and emotion. *Imagination, Cognition and Personality* 2:37-50.

Kunst-Wilson, W. R. & Zajonc, R. B. 1980. Affective discrimination of stimuli that cannot be recognized. *Science* 207:557-58.

Lane, R. D. Reiman, E. , Ahern, G. , Schwartz, G. E. , & Yun, L. 1998. Anterior cingulate cortex participates in the conscious experience of emotion. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Lane, R. 2000. Neural correlates of conscious emotional experience. In (R. Lane, L. Nadel, & G. Ahern, J. Allen, & A. Kaszniak, eds) *Cognitive Neuroscience of Emotion*. Oxford University Press.

Lane, R. D. , Ahern, G. L. , Schwartz, G. E. & Kaszniak, A. W. 1997. Is alexithymia the emotional equivalent of blindsight? *Biological Psychiatry* 42:834-44.

Ledoux, J. E. 1995. In search of an emotional system in the brain: Leaping from fear to emotion and consciousness. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Lewis, M. 1997. The self in self-conscious emotions. In (J. Snodgrass, R. Thompson, eds) *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.

Lishman, W. A. 1971. Emotion, consciousness, and will after brain bisection in man. *Cortex* 7:181-92.

Matsumoto, D. & Lee, M. 1993. Consciousness, volition, and the neuropsychology of facial expressions of emotion. *Consciousness and Cognition* 2:237-54.

Morris, J. S. , Ohman, A. , & Dolan, R. J. 1998. Conscious and unconscious emotional learning in the human amygdala. *Nature* 393:467-470.

Niedenthal, P. M. 1990. Implicit perception of affective information. *Journal of Experimental Social Psychology* 26:505-27.

Ochsner, K. 2000. Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology: General* 129:242-261.

Ohman, A. , Flykt, A. , & Lundqvist, D. 2000. Unconscious emotion: Evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. In (R. Lane, L. Nadel, & G. Ahern, eds) *Cognitive Neuroscience of Emotion*. Oxford University Press.

Peper, M. 2000. Awareness of emotions: A neuropsychological perspective. In (R. Ellis, ed) *The Caldron of Consciousness: Motivation, Affect and Self-organization*. John Benjamins Publishing Company.

Pratto, F. 1994. Consciousness and automatic evaluation. In (P. Niedenthal & S. Kitayama, eds) *The Heart's Eye: Emotional Influences in Perception and Attention*. Academic Press.

- Reiman, E. M. , Lane, R. D. , Ahern, G. L. , Schwartz, G. E. 1996. Positron emission tomography, emotion, and consciousness. In (S. Hamreoff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Rolls, E. T. 1995. A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In (M. Gazzaniga, eds) *The Cognitive Neurosciences*. MIT Press.
- Safran, J. D. & Greenberg, L. S. 1987. Affect and the unconscious: A cognitive perspective. In (R. Stern, ed) *Theories of the Unconscious and Theories of the Self*. Analytic Press.
- Wakefield, J. C. 1991. Why emotions can't be unconscious: An exploration of Freud's essentialism. *Psychoanalysis and Contemporary Thought* 14:29-67.
- Weiskrantz, L. 2000. Blindsight: Implications for the conscious experience of emotion. In (R. Lane, L. Nadel, & G. Ahern, eds) *Cognitive Neuroscience of Emotion*. Oxford University Press.
- Zajonc, R. 2000. Feeling and thinking: Closing the debate over the independence of affect. In (J. Forgas, ed) *Feeling and Thinking: The Role of Affect in Social Cognition*. Cambridge University Press.

6.21

Consciousness, Sleep, and Dreaming

- Arden, J. B. 1996. *Consciousness, Dreams, and Self: A Transdisciplinary Approach*. Psychosocial Press.
- Bentley, E. 2000. *Awareness: Biorhythms, Sleep and Dreaming*. Routledge.
- Bosinelli, M. 1995. Mind and consciousness during sleep. *Behavioural Brain Research* 69:195-201.
- Broughton, R. J. 1982. Human consciousness and sleep/waking rhythms: A review and some neuropsychological considerations. *Journal of Clinical Neuropsychology* 4:193-218.
- Combs, A. & Krippner, S. 1998. Dream sleep and waking reality: A dynamical view. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Flanagan, O. 1997. Prospects for a unified theory of consciousness or, what dreams are made of. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Foulkes, D. 1990. Dreaming and consciousness. *European Journal of Cognitive Psychology* 2:39-55.

- Gackenbach, J. & LaBerge, S. 1988. *Conscious Mind, Sleeping Brain: Perspectives on Lucid Dreaming*. Plenum Press.
- Green, C. & McGreery, C. 1994. *Lucid Dreaming: The Paradox of Consciousness During Sleep*. Routledge.
- Hearne, K. M. 1992. Prolucid dreaming, lucid dreams, and consciousness. *Journal of Mental Imagery* 16:119-123.
- Hobson, J. A. & Strickgold, R. 1995. The conscious state paradigm: A neurocognitive approach to waking, sleeping, and dreaming. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Hobson, J. A. 1998. The conscious state paradigm: A neuropsychological analysis of waking, sleeping, and dreaming. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Hobson, J. , Pace-Schott, E. , & Stickgold, R. 2000. Consciousness: Its vicissitudes in waking and sleep. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.
- Jones, B. E. 1998. The neural basis of consciousness across the sleep-waking cycle. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.
- Kahan, T. L. & LaBerge, S. 1994. Lucid dreaming as metacognition: Implications for cognitive science. *Consciousness & Cognition* 3:246-264.
- Kahan, T. L. & LaBerge, S. 1996. Cognition and metacognition in dreaming and waking: Comparisons of first and third-person ratings. *Dreaming* 6:235-249.
- Kahn, D. , Pace-Schott, E. F. & Hobson, J. A. 1997. Consciousness in waking and dreaming: The roles of neuronal oscillation and neuromodulation in determining similarities and differences. *Neuroscience* 78:13-38.
- Khan, D. , Krippner, S. , & Combs, A. 2000. Dreaming and the self-organizing brain. *Journal of Consciousness Studies* 7:4-11.
- King, C. D. 1947. Dream and the problem of consciousness. *Journal of General Psychology* 37:15-24.
- Kleitman, N. 1957. Sleep, wakefulness, and consciousness. *Psychological Bulletin* 54:354-359.
- LaBerge, S. 1985. *Lucid Dreaming*. J. P. Tarcher.

- LaBerge, S. , Levitan, L. , & Dement, W. C. 1986. Lucid dreaming: Physiological correlates of consciousness during REM sleep. *Journal of Mind and Behavior* 7:251-258.
- LaBerge, S. 1990. Lucid dreaming: Psychophysiological studies of consciousness during REM sleep. In (R. Bootsen, J. Kihlstrom, & D. Schacter, eds) *Sleep and Cognition*. American Psychological Association Press.
- LaBerge, S. 1998. Dreaming and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- LaBerge, S. & DeGracia, D. 2000. Varieties of lucid dreaming experience. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.
- Lindsley, D. B. 1960. Attention, consciousness, sleep, and wakefulness. In (H. Magoun & V. Hall, eds) *Handbook of Physiology. Section I: Neurophysiology*. American Physiological Society.
- Makeig, S. , Jung, T. , & Sejnowski, T. 2000. Awareness during drowsiness: Dynamics and electrophysiological correlates. *Canadian Journal of Experimental Psychology* 54:266-273.
- Monnier, M. 1952. Experimental work on sleep and other variations of consciousness. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Third Conference*. Josiah Macy Foundation.
- Moorcroft, W. & Breitenstein, J. 2000. Awareness of time during sleep. *Annals of Medicine* 32:236-238.
- Munglani, R. & Jones, J. G. 1992. Sleep and general anaesthesia as altered states of consciousness. *Journal of Psychopharmacology* 6:399-409.
- Pare, D. & Llinas, R. 1995. Conscious and pre-conscious processes as seen from the standpoint of sleep-waking cycle neurophysiology. *Neuropsychologia* 33:1155-1168.
- Revonsuo, A. 1995. Consciousness, dreams and virtual realities. *Philosophical Psychology* 8:35-58.
- Simon, C. W. & Emmons, W. 1956. EEG, consciousness, and sleep. *Science* 124:1066-1069.
- Stoyva, J. & Kamiya, J. 1968. Electrophysiological studies of dreaming as the prototype of a new strategy in the study of consciousness. *Psychological Review* 75:192-205.

6.2m

Consciousness in Psychoanalysis

- Balint, E. 1987. Memory and consciousness. *International Journal of Psychoanalysis* 68:475-483.
- Barr, H. & Langs, R. 1972. The psychoanalytic theory of consciousness. In *LSD: Personality and experience*. Wiley-Interscience.
- Bouveresse, J. 1995. *Wittgenstein Reads Freud: The Myth of the Unconscious*. Princeton University Press.
- Bower, G. H. 1990. Awareness, the unconscious, and repression: An experimental psychologist's perspective. In (J. Singer, ed) *Repression and Dissociation*. University of Chicago Press.
- Brakel L. W. 1989. Negative hallucinations, other irretrievable experiences and two functions of consciousness. *International Journal of Psychoanalysis* 70:461-89.
- Burston, D. 1986. The cognitive and dynamic unconscious: A critical and historical perspective. *Contemporary Psychoanalysis* 22:133-57.
- Chang, S. C. 1978. The psychology of consciousness. *American Journal of Psychotherapy* 32:105-116.
- Curtis, R. 1992. A process view of consciousness and the "self": Integrating a sense of connectedness with a sense of agency. *Psychological Inquiry*, 3:29-32.
- Eagle, M. N. 1987. The psychoanalytic and the cognitive unconscious. In (R. Stern, ed) *Theories of the Unconscious and Theories of the Self*. Analytic Press.
- Epstein, S. 1994. Integration of the cognitive and the psychodynamic unconscious. *American Psychologist* 49:409-24.
- Erdelyi, M. H. 1988. Issues in the study of unconscious and defense processes. In (M. Horowitz, ed) *Psychodynamics and Cognition*. University of Chicago Press.
- Foulkes, D. 1964. Theories of dream formation and recent studies of sleep consciousness. *Psychological Bulletin* 62:236-47.
- Globus, G. G. 1974. The problem of consciousness. *Psychoanalysis and Contemporary Science* 3:40-69.
- Haldane, J. 1988. Psychoanalysis, cognitive psychology and self-consciousness. In (P. Clark & C. Wright, eds) *Mind, Psychoanalysis and Science*. Blackwell.
- Herzog, P. S. 1991. *Conscious and Unconscious: Freud's Dynamic Distinction Reconsidered*. International Universities Press.

- Horowitz, M. J. & Stinson, C. H. 1995. Consciousness and processes of control. *Journal of Psychotherapy Practice and Research* 4:123-139.
- Joseph E. D. 1987. The consciousness of being conscious. *Journal of the American Psychoanalytic Association* 35:5-22.
- Klein, G. 1959. Consciousness in psychoanalytic theory. *Journal of the American Psychoanalytic Association* 7:5-34.
- Kubie, L. S. 1954. Psychiatric and psychoanalytic considerations of the problem of consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.
- Levy, D. 1996. *Freud among the Philosophers: The Psychoanalytic Unconscious and its Philosophical Critics*. Yale University Press.
- Masek, R. 1989. The overlooked problem of consciousness in psychoanalysis: Pierre Janet revisited. *Humanistic Psychologist* 17:274-279.
- MacIntyre, A. C. 1958. *The Unconscious: A Conceptual Study*. London.
- Miller, L. 1997. Freud and consciousness: The first one hundred years of neuropsychodynamics in theory and clinical practice. *Seminars in Neurology* 17:171-77.
- Moraglia, G. 1991. The unconscious in information processing and analytical psychology. *Journal of Analytical Psychology* 36:27-36.
- Natsoulas, T. 1992. Toward an improved understanding of Sigmund Freud's conception of consciousness. *Journal of Mind & Behavior* 13:171-92.
- Natsoulas, T. 1984-1996. Freud and consciousness I-XI. *Psychoanalysis and Contemporary Thought* 7:195-232, 8:183-220, 12:97-123, 12:619-62, 14:69-108, 15:305-48, 16:67-101, 16:597-631, 19:461-94.
- Natsoulas, T. 1995. A rediscovery of Sigmund Freud. *Consciousness and Cognition* 4:300-322.
- Olds, D. D. 1992. Consciousness: A brain-centered, informational approach. *Psychoanalytic Inquiry* 12:419-44.
- Opatow, B. 1997. The real unconscious: Psychoanalysis as a theory of consciousness. *Journal of the American Psychoanalytic Association* 45:865-90.
- Orbach, I. 1995. *The Hidden Mind: Psychology, Psychotherapy, and Unconscious Processes*. Wiley.

- Piaget, J. 1973. The affective unconscious and the cognitive unconscious. *Journal of the American Psychoanalytic Association* 21:249-261.
- Power, M. J. & Brewin, C. R. 1991. From Freud to cognitive science: A contemporary account of the unconscious. *British Journal of Clinical Psychology* 30:289-310.
- Power, M. J. 1997. Conscious and unconscious representations of meaning. In (M. Power & C. Brewin, eds) *The Transformation of Meaning in Psychological Therapies: Integrating Theory and Practice*. John Wiley.
- Rapaport, D. 1951. Consciousness: A Psychopathological and psychodynamic view. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Second Conference*. Josiah Macy Foundation.
- Rosenblatt A. D. & Thickstun J. T. 1994. Intuition and consciousness. *Psychoanalytic Quarterly* 63:696-714.
- Rubinfine, D. L. 1973. Notes toward a theory of consciousness. *International Journal of Psychoanalytic Psychotherapy* 2:391-410.
- Ryle, A. 1994. Consciousness and psychotherapy. *British Journal of Medical Psychology* 67:115-23.
- Schimek, J. G. 1975. A critical re-examination of Freud's concept of unconscious mental representation. *International Review of Psychoanalysis* 2:171-87.
- Shevrin, H. 1990. Subliminal perception and repression. In (J. Singer, ed) *Repression and Dissociation: Implications for Personality Theory, Psychopathology, and Health*. University of Chicago Press.
- Shevrin, H. 1992. The Freudian unconscious and the cognitive unconscious: Identical or fraternal twins? In (J. Barron, M. Eagle, & D. Wolitzky, eds) *Interface of Psychoanalysis and Psychology*. American Psychological Association.
- Shevrin, H. , Williams, W. J. , Marshall, R. E. , & Brakel, L. A. 1992. Event-related potential indicators of the dynamic unconscious. *Consciousness and Cognition* 1:340-66.
- Shevrin, H. , Bond, J. , Brakel, L. , Hertel, R. & Williams, W. 1996. *Conscious and Unconscious Processes: Psychodynamic, Cognitive, and Neurophysiological Convergences*. Guilford Press.
- Shevrin, H. 1998. The Freud-Rapaport theory of consciousness. In (R. Bornstein & J. Masling, eds) *Empirical Perspectives on the Psychoanalytic Unconscious*. American Psychological Association.
- Slipp, S. 2000. Subliminal stimulation research and its implications for psychoanalytic theory and

treatment. *Journal of the American Academy of Psychoanalysis* 28:305-320.

Smith, D. 2000. Freudian science of consciousness: Then and now. *Neuro-psychoanalysis* 2:38-45.

Solomon, R. C. 1974. Freud and "unconscious motivation". *Journal for the Theory of Social Behaviour* 4:191-216.

Solms, M. 1997. What is consciousness? *Journal of the American Psychoanalytic Association* 45:681-703.

Spence, D. P. & Holland, B. 1962. The restricting effects of awareness: A paradox and an explanation. *Journal of Abnormal and Social Psychology* 64:163-74.

Stein, D. J. (ed) 1997. *Cognitive Science and the Unconscious*. American Psychiatric Press.

Strauss, A. 1955. Unconscious mental processes and the psychosomatic concept. *International Journal of Psychoanalysis* 36:307-19.

van der Waals, E. G. 1949. The psycho-analytical and the phenomenological concept of consciousness. *International Journal of Psychoanalysis* 30:207.

Wakefield, J. C. 1990. Why instinctual impulses can't be unconscious: An exploration of Freud's cognitivism. *Psychoanalysis and Contemporary Thought* 13:265-88.

Wakefield, J. C. 1991. Why emotions can't be unconscious: An exploration of Freud's essentialism. *Psychoanalysis and Contemporary Thought* 14:29-67.

Weinberger, J. & Weiss, J. 1997. Psychoanalytic and cognitive conceptions of the unconscious. In (D. Stein, ed) *Cognitive Science and the Unconscious*. American Psychiatric Press.

Westen, D. 1992. The cognitive self and the psychoanalytic self: Can we put our selves together? *Psychological Inquiry* 3:1-13.

Woody, J. M. & Phillips, J. 1995. Freud's project for a scientific psychology after 100 years: The unconscious mind in the era of cognitive neuroscience. *Philosophy, Psychiatry, and Psychology* 2:123-34.

Zilboorg, G. 1951. Variations in the scope of awareness. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Second Conference*. Josiah Macy Foundation.

6.2n

Consciousness and Time

Allport, D. A. 1968. Phenomenal similarity and the perceptual moment hypothesis. *British Journal of Psychology* 59:395-406.

Banks, R. & Cappon, D. 1962. Effect of reduced sensory input on time perception. *Perceptual and Motor Skills* 14:74.

Block, R. A. 1979. Time and consciousness. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 1*. Academic Press.

Block, R. A. (ed) 1990. *Cognitive Models of Psychological Time*. Lawrence Erlbaum.

Block, R. A. 1996. Psychological time and memory systems of the brain. In (J. Fraser & M. Soulsby, eds) *Dimensions of Time and Life: The Study of Time*, volume 8. International Universities Press.

Brown, J. W. 1990. Psychology of time awareness. *Brain and Cognition* 14:144-64.

Brown, J. W. 1991. *Self and Process: Brain States and the Conscious Present*. Springer-Verlag.

Brown, J. 2000. *Mind and Nature: Essays on Time and Subjectivity*. Whurr Publishers.

Cohen, J. 1954. The experience of time. *Acta Psychologica* 10:207-19.

Dennett, D. C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15:183-201.

Eisler, H. 1975. Subjective duration and psychophysics. *Psychological Review* 82:429-50.

Fraser, J. T. (ed) 1989. *Time and Mind: Interdisciplinary Issues*. International Universities Press.

Goody, W. 1967. Introduction to problems of time awareness. *Studium Generale* 20:33-41.

Hicks, R. E. , Miller, G. W. , Gaes, G. , & Bierman, K. 1977. Concurrent processing demands and the experience of time-in-passing. *American Journal of Psychology* 90:431-46.

Hoagland, H. 1950. Consciousness and the chemistry of time. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the First Conference*. Josiah Macy Foundation.

Hoagland, H. 1943. The chemistry of time. *Scientific Monthly* 56:56-61.

- Knight, R. T. & Grabowecky, M. 1995. Escape from linear time: Prefrontal cortex and conscious experience. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Melges, F. T. 1989. Disorders of time and the brain in severe mental illness. In (J. Fraser, ed) *Time and Mind: Interdisciplinary Issues*. International Universities Press.
- Michon, J. A. 1975. Time experience and memory processes. In (J. Fraser & N. Lawrence, eds) *The Study of Time II*. Springer-Verlag.
- Michon, J. A. 1972. Processing of temporal information and the cognitive theory of time experience. In (J. Fraser, F. Haber, & G. Muller, eds) *The Study of Time*. Springer-Verlag.
- Moiseeva, N. I. 19xx. Perception of time by human consciousness. *Chronobiologia* 15:301-317.
- Natsoulas, T. 1993. The stream of consciousness: II. William James's specious present. *Imagination, Cognition and Personality* 12:367-385.
- Newman, M. A. 1982. Time as an index of expanding consciousness with age. *Nursing Research* 31:290-293.
- Orme, J. E. 1969. *Time, Experience and Behaviour*. Illife.
- Ornstein, R. E. 1969. *On the Experience of Time*. Harmondsworth.
- Poppel, E. 1988. *Mindworks: Time and Conscious Experience*. Harcourt Brace Jovanovich.
- Poppel, E. & Schwender, D. 1993. Temporal mechanisms of consciousness. *International Anesthesiology Clinics* 31:27-38.
- Proust, J. 1994. Time and conscious experience. In (C. C. Gould, ed) *Artifacts, Representations, and Social Practice*. Kluwer.
- Reidhead, V. A. & Wolford, J. B. 1998. Context, conditioning, and meaning of time-consciousness in a Trappist monastery. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Robertson, T. B. 1923. Consciousness and the sense of time. *Scientific Monthly* 16:649-657.
- Sanders, S. A. 1986. Development of a tool to measure subjective time experience. *Nursing Research* 35:178-182.

Schaltenbrand, G. 1967. Consciousness and time. *Annals of the New York Academy of Sciences* 138:632-645.

Strong, C. A. 1896. Consciousness and time. *Psychological Review* 3:149-57.

Stroud, J. M. 1967. The fine structure of psychological time. *Annals of the New York Academy of Sciences* 138:623-631.

Stroud, J. M. 1957. The fine structure of psychological time. In (H. Quastler, ed) *Information Theory in Psychology: Problems and Methods*. Free Press.

6.2o

Self-Consciousness

Asendorpf, J. B. , Warkentin, V. , & Baudonniere, P. 1996. Self-awareness and other-awareness. II: Mirror self-recognition, social contingency awareness, and synchronic imitation. *Developmental Psychology* 32q:313-321.

Butterworth, G. 1995. The self as an object of consciousness in infancy. In (P. Rochat, ed) *The Self in Infancy: Theory and Research*. Elsevier.

Carver, C. & Scheier, M. F. 1983. Self-awareness and the self-regulation of behaviour. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Cheeks, J. M. & Briggs, S. R. 1982. Self-consciousness and aspects of personality. *Journal of Research in Personality* 16:401-8.

Cooney, B. 1979. The neural basis of self-consciousness. *Nature and System* 1:16-31.

Duval, S. & Wicklund, R. A. 1972. *A Theory of Objective Self-Awareness*. Academic Press.

Dymond, S. & Barnes, D. 1997. Behavior-analytic approaches to self-awareness. *Psychological Record* 47:181-200.

Feinberg, T. E. 1997. Some interesting perturbations of the self in neurology. *Seminars in Neurology* 17:129-35.

Fenigstein, A. , Scheier, M. F. & Buss. A. H. 1975. Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology* 43:522-27.

- Fenigstein, A. 1997. Self-consciousness and its relation to psychological mindedness. In (M. McCallum & W. Piper, eds) *Psychological Mindedness: A Contemporary Understanding*. Lawrence Erlbaum.
- Frith, C. D. 1996. The role of the prefrontal cortex in self-consciousness: the case of auditory hallucinations. *Philosophical Transactions of the Royal Society of London* B351:1505-12.
- Freeman, W. & Watts, J. W. 1941. The frontal lobes and consciousness of self. *Psychosomatic Medicine* 3:111-19.
- Gallup, G. G. 1998. Self-awareness and the evolution of social intelligence. *Behavioural Processes* 42:239-247.
- Gardiner, J. 2000. On the objectivity of subjective experiences and auto-noetic and noetic consciousness. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.
- Hart, D. & Fegley, S. 1997. The development of self-awareness and self-understanding in cultural context. In (U. Neisser & D. Jopling, eds) *The Conceptual Self in Context*. Cambridge University Press.
- Hart, D. & Fegley, S. 1994. Social imitation and the emergence of a mental model of self. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.
- James, W. 1890. The consciousness of self. In *The Principles of Psychology*.
- Jaynes, J. 1976. *The Origins of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin.
- Kessel, F. S, Cole, P. M. & Johnson, D. L. (eds) *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.
- Kihlstrom, J. F. 1997. Consciousness and me-ness. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Kihlstrom, J. F. & Klein, S. B. 1997. Self-knowledge and self-awareness. In (J. Snodgrass, R. Thompson, eds) *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.
- Kinsbourne, M. 1995. Awareness of one's own body: An attentional theory of its nature, development, and brain basis. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Kinsbourne, M. 1998. Representations in consciousness and the neuropsychology of insight. In (X.

Amador & A. David, eds) *Insight and Psychosis*. Oxford University Press.

Kunzendorf, R. G. 1988. Self-consciousness as the monitoring of cognitive states: A theoretical perspective. *Imagination, Cognition and Personality* 7:3-22.

Kunzendorf, R. G. , Beltz, S. M. & Tymowicz, G. 1992. Self-awareness in autistic subjects and deeply hypnotized subjects: Dissociation of self-concept versus self-consciousness. *Imagination, Cognition and Personality* 11:129-41.

Kunzendorf, R. 2000. Individual differences in self-conscious source monitoring: Theoretical, experimental, and clinical considerations. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Levine, B. 2000. Self-regulation and autonoetic consciousness. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Lewis, M. 1991. Ways of knowing: Objective self-awareness or consciousness. *Developmental Review* 11:231-43.

Lewis, M. 1994. Myself and me. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Markova, I. 1990. The development of self-consciousness: Baldwin, Mead, and Vygotsky. In (J. Faulconer & R. Williams, eds) *Reconsidering Psychology*. Duquesne University Press.

Mitchell, R. W. 1993. Mental models of mirror self-recognition: Two theories. *New Ideas in Psychology* 11:295-325.

Mitchell, R. W. 1994. Multiplicities of self. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Mitchell, R. W. 1997. A comparison of the self-awareness and kinesthetic-visual matching theories of self-recognition: Autistic children and others. In (J. Snodgrass, R. Thompson, eds) *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.

Mollon, P. 1987. Self-awareness, self-consciousness, and preoccupation with self. In (K. Yardley & T. Honess, eds) *Self and Identity: Psychosocial Perspectives*. Wiley.

Morin, A. & Everett, J. 1990. Inner speech as a mediator of self-awareness, self-consciousness, and self-knowledge: An hypothesis. *New Ideas in Psychology* 8:337-56.

Nasby, W. 1989. Private self-consciousness, self-awareness, and the reliability of self-reports. *Journal of*

Personality and Social Psychology 56:950-7.

Neisser, U. 1991. Five kinds of self-knowledge. *Philosophical Psychology* 1:35-59.

Neisser, U. 1992. The development of consciousness and the acquisition of self. In (F. Kessel, P. Cole, & D. L. Johnson, eds) *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.

Parker, S. T. , Mitchell, R. M. , & Boccia, M. L. 1994. *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Sass, L. 2000. Schizophrenia, self-experience, and the so-called "negative symptoms": Reflections on hyperreflexivity. In (D. Zahavi, ed) *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience*. John Benjamins.

Shotter, J. 1983. Consciousness and self-consciousness: Inner games and alternative realities. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Shrauger, J. S. & Osberg, T. M. 1983. Self-awareness: The ability to predict one's subsequent behaviour. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Siegrist, M. 1995. Inner speech as a cognitive process mediating self-consciousness and inhibiting self-deception. *Psychological Reports* 76:259-65.

Snodgrass, J. G. & Thompson, R. L. (eds) 1997. *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.

Titchener, E. B. 1911. A note on the consciousness of self. *American Journal of Psychology* 22:540-52.

Watson, J. S. 1994. Detection of self: The perfect algorithm. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Watson, P. J. , Morris, R. J. , Ramsey, A. Hickman, S. E. 1996. Further contrasts between self-reflectiveness and internal state awareness factors of private self-consciousness. *Journal of Psychology* 130:183-92.

6.2p

Development of Consciousness

Anderson, J. R. 1984. The development of self-recognition: A review. *Developmental Psychobiology*

17:35-49.

Brainerd, C. J. , Stein, L. M. , & Reyna, V. F. 1998. On the development of conscious and unconscious memory. *Developmental Psychology* 34:342-357.

Briskin, A. S. 1974. A developmental model of self-awareness. *Counseling and Values* 18:79-85.

Burgess, J. A. & Tawia, S. A. 1996. When did you first begin to feel it? Locating the beginnings of human consciousness? *Bioethics* 10:1-26.

Butterworth, G. 1995. The self as an object of consciousness in infancy. In (P. Rochat, ed) *The Self in Infancy: Theory and Research*. Elsevier.

Davis, L. H. 1989. Self-consciousness in chimps and pigeons. *Philosophical Psychology* 2:249-59.

Flavell, J. H. 1993. Young children's understanding of thinking and consciousness. *Current Directions in Psychological Science* 2:40-43.

Flavell, J. H. , Green, F. L. , & Flavell, E. R. 1993. Children's understanding of the stream of consciousness. *Child Development* 64:387-398.

Flavell, J. H. , Green, F. L. , Flavell, E. R. & Grossman, J. B. 1997. The development of children's knowledge about inner speech. *Child Development* 68:39-47.

Flavell, J. H. , Green, F. L. , Flavell, E. R. 1995. The development of children's knowledge about attentional focus. *Developmental Psychology* 31:706-12.

Flavell, J. , Green, F. , & Flavell, E. 2000. Development of children's awareness of their own thoughts. *Journal of Cognition & Development* 1:97-112.

Foulkes, D. 1999. *Children's Dreaming and the Development of Consciousness*. Harvard University Press.

Gallagher, S. & Meltzoff, A. 1996. The earliest sense of self and others: Merleau-Ponty and recent developmental studies. *Philosophical Psychology* 9:211-33.

Gopnik, A. & Meltzoff, A. N. 1994. Minds, bodies, and persons: Young children's understanding of the self and others as reflected in imitation and theory of mind research. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Griffin, S. 1991. Young children's awareness of their inner world: A neo-structural analysis of the

- development of intrapersonal intelligence. In (R. Case, ed) *The Mind's Staircase: Exploring the Conceptual Underpinnings of Children's Thought and Knowledge*. Lawrence Erlbaum.
- Kagan, J. 1981. *The Second Year: The Emergence of Self-Awareness*. Harvard University Press.
- Kuhn, D. 2000. Metacognitive development. *Current Directions in Psychological Science* 9:178-181.
- Lewis, M. 1990. The development of intentionality and the role of consciousness. *Psychological Inquiry* 1:231-247.
- Lewis, M. 1991. Ways of knowing: Objective self-awareness or consciousness. *Developmental Review* 11:231-43.
- Lunzer, E. A. 1979. The development of consciousness. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness*. Academic Press.
- Marbach, E. 1987. Laws of consciousness as norms of mental development. In (B. Inhelder, D. de Caprona, & A. Cornu-Wells, eds) *Piaget Today*. Lawrence Erlbaum.
- Markova, I. 1990. The development of self-consciousness: Baldwin, Mead, and Vygotsky. In (J. Faulconer & R. Williams, eds) *Reconsidering Psychology*. Duquesne University Press.
- McCune, L. 1993. The development of play as the development of consciousness. In (M. Bornstein & A. O'Reilly, eds) *The Role of Play in the Development of Thought*. Jossey-Bass.
- Mounoud, P. 1990. Consciousness as a necessary transitional phenomenon in cognitive development. *Psychological Inquiry* 1:253-58.
- Neisser, U. 1992. The development of consciousness and the acquisition of skill. In (F. Kessel, P. Cole, & D. Johnson, eds) *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.
- Parker, S. T. , Mitchell, R. M. , & Boccia, M. L. 1994. *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.
- Piaget, J. 1954. The problem of consciousness in child psychology: Developmental changes in awareness. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Fourth Conference*. Josiah Macy Foundation.
- van Eenwyk, J. R. 1996. Chaotic dynamics and the development of consciousness. In (E. MacCormac & M. Stamenov, eds) *Fractals of Brain, Fractals of Mind: In Search of a Symmetry Bond*. John Benjamins.

- Wheeler, M. 2000. Varieties of consciousness and memory in the developing child. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.
- Wilber, K. 1979. A developmental view of consciousness. *Journal of Transpersonal Psychology* 11:1-21.
- Zelazo, P. D. 1996. Towards a characterization of minimal consciousness. *New Ideas in Psychology* 14:63-80.
- Zelazo, P. R. & Zelazo, P. D. 1998. The emergence of consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.
- Zelazo, P. 2000. Self-reflection and the development of consciously controlled processing. In (P. Mitchell & K. Riggs, eds) *Children's Reasoning and the Mind*. Psychology Press/Taylor & Francis.

6.2q

The Stream of Consciousness

- Antrobus, J. S. , Singer, J. L. , & Greenberg, S. 1966. Studies in the stream of consciousness: Experimental enhancement and suppression of spontaneous cognitive processes. *Perceptual and Motor Skills* 23:399-417.
- Baars, B. J. 1993. How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity? In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174).
- Bakan, P. 1978. Two streams of consciousness: A typological approach. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.
- Bonanno, G. A. & Singer, J. L. 1993. Controlling one's stream of thought through perceptual and reflective processing. In (D. Wegner & J. Pennebaker, eds) *Handbook of Mental Control*. Prentice-Hall.
- Capek, M. 1950. Stream of consciousness and "duree reelle." *Philosophy and Phenomenological Research* 10:331-353.
- Diaz, J. 1996. The stream revisited: A process model of phenomenological consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Dainton, B. 2000. *Stream of Consciousness: Unity and continuity in conscious experience*. Routledge.
- Flanagan, O. 1992. The stream of consciousness. In *Consciousness Reconsidered*. MIT Press.

Flavell, J. H. , Green, F. L. , & Flavell, E. R. 1993. Children's understanding of the stream of consciousness. *Child Development* 64:387-398.

Gurwitsch, A. 1943. William James' theory of the "transitive parts" of the stream of consciousness. *Philosophy and Phenomenological Research* 3:449-477.

James, W. 1990. The stream of thought. In *Principles of Psychology*.

James, W. 1992. The stream of consciousness. In *Psychology: The Briefer Course*.

Klinger, E. 1978. Modes of normal conscious flow. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Mueller, E. T. 1990. *Daydreaming in Humans and Machines: A Computer Model of the Stream of Thought*. Ablex.

Natsoulas, T. 1987. The six basic concepts of consciousness and William James' stream of thought. *Imagination, Cognition, and Personality* 6:289-319.

Natsoulas, T. 1988. Sympathy, empathy, and the stream of consciousness. *Journal for the Theory of Social Behaviour* 18:169-195.

Natsoulas, T. 1992-1996. The stream of consciousness: Parts I-XVI. *Imagination, Cognition, and Personality* 12:3-21, 12:367-85, 13:73-90, 13:229-47, 13:347-66, 14:59-77, 14:131-49, 14:333-52, 15:171-91, 16:161-80, 16:281-300, 17:45-64, 17:123-40, 17:229-47.

Penfield, W. 1955. The permanent record of the stream of consciousness. *Acta Psychologica* 11:47-69.

Petchkovsky, L. 2000. 'Stream of consciousness' and 'ownership of thought' in indigenous people in Central Australia. *Journal of Analytical Psychology* 45:577-597.

Pollio, H. R. 1990. The stream of consciousness since James. In (M. Johnson & T. Henley, eds) *Reflections on "The Principles of Psychology": William James after a Century*. Lawrence Erlbaum.

Pope, K. S. & Singer, J. L. (eds) 1978. *The Stream of Consciousness: Scientific Investigations into the Flow of Human Experience*. Plenum Press.

Pope, K. S. 1978. How gender, solitude, and posture influence the stream of consciousness. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

- Rychlak, J. F. 1978. The stream of consciousness: Implications for a humanistic psychological theory. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.
- Schuetz, A. 1940. William James's concept of the stream of thought, phenomenologically interpreted. *Journal of Philosophy* 37:673-74.
- Singer, J. L. 1974. Daydreaming and the stream of thought. *American Scientist* 62:417-425.
- Singer, J. L. 1975. Navigating the stream of consciousness: Research in daydreaming and related inner experience. *American Psychologist* 30:727-738.
- Singer, J. L. 1978. Experimental studies of daydreaming and the stream of thought. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.
- Singer, J. L. 1998. Daydreams, the stream of consciousness, and self-representations. In (R. Bornstein & J. Masling, eds) *Empirical Perspectives on the Psychoanalytic Unconscious*. American Psychological Association.
- Strange, J. R. 1978. A search for the sources of the stream of consciousness. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

6.2r

Foundational Issues

- Baars, B. J. 1986. What is a theory of consciousness a theory of? The search for criterial constraints on theory. *Imagination, Cognition, and Personality* 1:3-24.
- Baars, B. J. 1994. A thoroughly empirical approach to consciousness. *Psyche* 1.
- Baars, B. J. 1996. Understanding subjectivity: Global workspace theory and the resurrection of the observing self. *Journal of Consciousness Studies* 3:211-17.
- Benoit, P. J. & Benoit, W. L. 1986. Consciousness: The mindlessness/mindfulness and verbal report controversies. *Western Journal of Speech Communication* 50:41-63.
- Bindra, D. 1970. The problem of subjective experience. *Psychological Review* 77:581-84.
- Blanshard, B. & Skinner, B. F. 1966. The problem of consciousness: A debate. *Philosophy and Phenomenological Research* 27:317-37.

- Carlson, R. A. 1992. Starting with consciousness. *American Journal of Psychology* 105:598-604.
- Casler L. 1976. The "consciousness problem" is not the problem. *Perceptual and Motor Skills* 42:227-32.
- Hebb, D. O. 1954. The problem of consciousness and introspection. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.
- Kihlstrom, J. F. 1987. What this discipline needs is a good ten-cent taxonomy of consciousness. *Canadian Psychology* 28:116-118.
- Merikle, P. M. 1984. Toward a definition of awareness. *Bulletin of the Psychonomic Society* 22:449-50.
- Natsoulas, T. 1974. The subjective, experiential element in perception. *Psychological Bulletin* 81:611-31.
- Natsoulas, T. 1981. Basic problems of consciousness. *Journal of Personality and Social Psychology* 41:132-78.
- Natsoulas, T. 1990. Is consciousness what psychologists actually examine? *American Journal of Psychology* 105:363-84.
- Revonsuo, A. 1993. Is there a ghost in the cognitive machinery? *Philosophical Psychology* 6:387-405.
- Rychlak, J. F. 1997. *In Defense of Human Consciousness*. American Psychological Association.
- Tolman, E. C. 1935. Psychology versus immediate experience. *Philosophy of Science* 2:356-80.
- Wilson, D. L. 1978. Brain mechanisms, consciousness, and introspection. In (A. Sugarman & R. Tarter, eds) *Expanding Dimensions of Consciousness*. Springer.
- Zener, K. 1952. Significance of the experience of the individual for the science of psychology. *Minnesota Studies in the Philosophy of Science* 2:354-69.

6.2s

Consciousness and Psychology, Misc

- Aurell, G. 1979. Perception: A model comprising two modes of consciousness. *Perceptual and Motor Skills* 49:431-44.
- Aurell, G. 1989. Man's triune conscious mind, parts I, II, and III. *Perceptual and Motor Skills* 68:747-54,

78:31-39, 81:463-66

Bolton, N. 1983. Forms of awareness. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Bowden, E. M. 1997. The effect of reportable and unreportable hints on anagram solution and the aha! experience. *Consciousness and Cognition* 6:545-573.

Davidson, J. M. & Davidson, R. J. (eds) 1980. *The Psychobiology of Consciousness*. Plenum.

Davidson, R. , Schwartz, G. & Shapiro, D. (eds) 1983. *Consciousness and Self-Regulation*. Plenum.

Delacour, J. 1997. Object perception and recognition: A model for the scientific study of consciousness. *Theory and Psychology* 7:257-62.

Dulany, D. E. 1991. Conscious representation and thought systems. In (R. Wyer & T. Srull, eds) *The Content, Structure, and Operation of Thought Systems*. Lawrence Erlbaum.

Dulany, D. E. 1997. Consciousness in the explicit (deliberative) and implicit (evocative). In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Farthing, G. W. 1992. *The Psychology of Consciousness*. Prentice Hall.

Frith C. D. 1979. Consciousness, information processing and schizophrenia. *British Journal of Psychiatry* 134:225-35.

Greenberg, G. & Tobach, E. (eds) 1987. *Cognition, language, and consciousness: Integrative levels*. Lawrence Erlbaum.

Higgins, E. E. & Bargh, J. A. 1992. Unconscious sources of subjectivity and suffering: Is consciousness the solution? In (L. Martin & A. Tesser, eds) *The Construction of Social Judgments*. Lawrence Erlbaum

Hilgard, E. R. 1977. *Divided Consciousness: Multiple Controls in Human Thought and Action*. Wiley.

Hilgard, E. R. 1977. Controversies over consciousness and the rise of cognitive psychology. *Australian Psychologist* 12:7-27.

Hilgard E. R. 1977. The problem of divided consciousness: A neodissociation interpretation. *Annals of the New York Academy of Sciences* 296:48-59.

Hilgard, E. R. 1980. Consciousness in contemporary psychology. *Annual Review of Psychology* 31:1-26.

- Hilgard, E. R. 1992. Divided consciousness and dissociation. *Consciousness and Cognition* 1:16-31.
- Hirst, W. 1995. Cognitive aspects of consciousness. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Humphrey, N. 1992. *A History of the Mind*. Simon and Schuster.
- Jonassen, D. H. 1979. Video-mediated, objective self-awareness, self-perception, and locus of control. *Perceptual and Motor Skills* 48:255-265.
- Kihlstrom, J. F. 1993. The continuum of consciousness. *Consciousness and Cognition* 2:334-54.
- Lashley, K. S. 1923. The behavioristic interpretation of consciousness. *Psychological Review* 30:237-72.
- Lewicki, P. , Czyzewska, M. & Hill, T. 1997. Cognitive mechanisms for acquiring "experience": The dissociation between conscious and nonconscious cognition. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Mandler, G. 1975. Consciousness: respectable, useful, and probably necessary. In (R. Solso, ed) *Information Processing and Cognition*. Lawrence Erlbaum.
- Mandler, G. 1988. Problems and direction in the study of consciousness. In (M. Horowitz, ed) *Psychodynamics and Cognition*. University of Chicago Press.
- Mandler, G. 1992. Toward a theory of consciousness. In (H. G. Geissler, S. W. Link, & J. T. Townsend, eds) *Cognition, Information Processing, and Psychophysics: Basic Issues*. Lawrence Erlbaum.
- Mandler, G. 1997. Consciousness redux. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Mandler, G. & Nakamura, Y. 1987. Aspects of consciousness. *Personality and Social Psychology Bulletin* 13:299-313.
- Natsoulas, T. 1984. Personality and consciousness: A theoretical essay. *Cognition and Brain Theory* 7:135-66.
- Oakley, D. A. & Eames, L. C. 1986. The plurality of consciousness. In (D. Oakley, ed) *Mind and Brain*. Methuen.
- Oatley, K. 1988. On changing one's mind: A possible function of consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

- Ornstein, R. E. (ed) 1974. *The Nature of Human Consciousness: A Book of Readings*. Viking Press.
- Ornstein, R. E. 1977. *The Psychology of Consciousness*. Harcourt Brace Jovanovich.
- Posner, M. I. & Klein, M. 1973. On the functions of consciousness. In (S. Kornblum, ed) *Attention and Performance*, vol 4. Academic Press.
- Schwartz, G. & Shapiro, D. (eds) 1976. *Consciousness and Self-regulation*. Plenum.
- Schwartz, G. 2000. Individual differences in subtle awareness and levels of awareness: Olfaction as a model system. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.
- Schwarz, N. & Clore, G. L. 1996. Feelings and phenomenal experiences. In *Social Psychology: Handbook of Basic Principles*. Guilford Press.
- Shallice, T. 1991. The revival of consciousness in cognitive science. In (W. Kessen, A. Ortony, & F. Craik, eds) *Memories, Thoughts, and Emotions: Essays in Honor of George Mandler*. Lawrence Erlbaum.
- Solso, R. (ed) 1975. *Information Processing and Consciousness*. Lawrence Erlbaum.
- Sperry, R. W. 1987. Structure and significance of the consciousness revolution. *Journal of Mind & Behavior* 8:37-65.
- Sperry, R. W. 1995. The riddle of consciousness and the changing scientific worldview. *Journal of Humanistic Psychology* 35:7-33.
- Stout, M. 2001. *The Myth of Sanity: Divided Consciousness and the Promise of Awareness*. Viking/Penguin Books.
- Underwood, G. & Stevens, R. (eds) 1979. *Aspects of Consciousness: Volume 1, Psychological Issues*. Academic Press.
- Underwood, G. & Stevens, R. (eds) 1981. *Aspects of Consciousness: Volume 2, Structural Issues*. Academic Press.
- Underwood, G. & Stevens, R. (eds) 1982. *Aspects of Consciousness: Volume 3, Awareness and Self-Awareness*. Academic Press.
- Underwood, G. & Stevens, R. (eds) 1984. *Aspects of Consciousness: Volume 4, Clinical Issues*. Academic Press.

Underwood, G. (ed) 1996. *Implicit Cognition*. Oxford University Press.

von der Malsburg, C. 1986. Am I thinking assemblies? In (G. Palm & A. Aertsen, eds) *Brain Theory*. Springer.

von der Malsburg, C. 1997. The coherence definition of consciousness. In (M. Ito, Y. Miyashita, & E. T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.

Wallace, B. & Fisher, L. 2000. Biological rhythms and individual differences in consciousness. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Wilks, Y. 1984. Machines and consciousness. In (C. Hookway, ed) *Minds, Machines and Evolution*. Cambridge University Press.

6.3

Consciousness and Physics

6.3a

The Interpretation of Quantum Mechanics

Albert, D. & Loewer, A. 1988. Interpreting the many-worlds interpretation. *Synthese* 77:195-213.

Butterfield, J. 1996. Whither the minds? *British Journal for the Philosophy of Science* 47:200-??.

Butterfield, J. 1998. Quantum curiosities of psychophysics. In (J. Cornwell, ed) *Consciousness and Human Identity*. Oxford University Press.

Byrne, A. & Hall, N. 1999. Chalmers on consciousness and quantum mechanics. *Philosophy of Science* 66:370-90.

Goertzel, B. 1992. Quantum theory and consciousness. *Journal of Mind and Behavior* 13:29-36.

Goswami, A. 1989. The idealistic interpretation of quantum mechanics. *Physics Essays* 2:385-400.

Goswami, A. 1990. Consciousness in quantum physics and the mind-body problem. *Journal of Mind and Behavior* 11:75-96.

- Klein, S. 1991. The duality of psycho-physics. In (A. Gorea, ed) *Representations of Vision*. Cambridge University Press.
- Lehner, C. 1997. What it feels like to be in a superposition, and why: Consciousness and the interpretation of Everett's quantum mechanics. *Synthese* 110:191-216.
- Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford University Press.
- Lockwood, M. 1996. Many-minds interpretations of quantum mechanics. *British Journal for the Philosophy of Science* 47:159-88.
- Mulhauser, G. 1995. Materialism and the "problem" of quantum measurement. *Minds and Machines* 5:207-17.
- Mulhauser, G. 1995. On the end of a quantum-mechanical romance. *Psyche* 2(19).
- Page, D. N. 1995. Attaching theories of consciousness to Bohmian quantum mechanics. Manuscript.
- Page, D. N. 1996. Sensible quantum mechanics: Are probabilities only in the mind? *International Journal of Modern Physics D* 5:583-96.
- Penrose, R. 1987. Quantum physics and conscious thought. In (B. Hiley & D. Peat, eds) *Quantum Implications: Essays in Honour of David Bohm*. Methuen.
- Shanks, N. 1995. Minds, brains, and quantum mechanics. *Southern Journal of Philosophy* 33:243-60.
- Squires, E. J. 1991. One mind or many? A note on the Everett interpretation of quantum theory. *Synthese* 89:283-6.
- Squires, E. J. 1993. Quantum theory and the relation between the conscious mind and the physical world. *Synthese* 97:109-23.
- Squires, E. J. 1994. Quantum theory and the need for consciousness. *Journal of Consciousness Studies* 1:201-4.
- Squires, E. J. 1998. Why are quantum theorists interested in consciousness? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Stapp, H. P. 1991. Quantum propensities and the brain-mind connection. *Foundations of Physics* 21:1451-77.

Stapp, H. P. 1993. *Mind, Matter, and Quantum Mechanics*. Springer-Verlag.

Stapp, H. P. 1995. Why classical mechanics cannot accommodate consciousness but quantum mechanics can. *Psyche* 2(5).

Stapp, H. P. 1998. The evolution of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Wigner, E. 1961. Remarks on the mind-body problem. In (I. Good, ed) *The Scientist Speculates*. Heineman.

Woo, C. H. 1981. Consciousness and quantum interference: An experimental approach. *Foundations of Physics* 11:933-44.

6.3b

Quantum Mechanisms of Consciousness

Bass, L. 1975. A quantum-mechanical mind-body interaction. *Foundations of Physics* 5:159-72.

Beck, F. & Eccles, J. 1992. Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Science USA* 89:11357-61.

Beck, F. 1998. Synaptic transmission, quantum-state selection, and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Berezin, A. A. 1992. Correlated isotopic tunneling as a possible model for consciousness. *Journal of Theoretical Biology* 154:415-20.

Dyer, M. G. 1994. Quantum physics and consciousness, creativity, computers: A commentary on Goswami's quantum-based theory of consciousness and free will. *Journal of Mind and Behavior* 15:265-90.

Eccles, J. C. 1986. Do mental events cause neural events analogously to the probability fields of quantum mechanics? *Proceedings of the Royal Society of London B* 227:411-28.

Germine, M. 1991. Consciousness and synchronicity. *Medical Hypotheses* 36:277-83.

Globus, G. 1997. Nonlinear brain systems with nonlocal degrees of freedom. *Journal of Mind and Behavior*.

- Globus, G. 1998. Self, cognition, qualia, and world in quantum brain dynamics. *Journal of Consciousness Studies* 5:34-52.
- Grush, R. & Churchland, P. 1995. Gaps in Penrose's toiling. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.
- Hameroff, S. R. 1994. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1:91-118.
- Hameroff, S. R. & Penrose, R. 1996. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Hameroff, S. R. & Penrose, R. 1996. Conscious events as orchestrated space-time selections. *Journal of Consciousness Studies* 3:36-53. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.
- Hameroff, S. R. & Scott, A. 1998. A Sonoran afternoon: A dialogue on quantum mechanics and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Jibu, M. & Yasue, K. 1995. *Quantum Brain Dynamics and Consciousness: An Introduction*. John Benjamins.
- Jibu, M. & Yasue, K. 1997. Magic without magic: Meaning of quantum brain dynamics. *Journal of Mind and Behavior*.
- King, C. 1997. Chaos, quantum mechanics, and the conscious brain. *Journal of Mind and Behavior*.
- Lahav, R. & Shanks, N. 1992. How to be a scientifically respectable 'property dualist'. *Journal of Mind and Behavior* 13:211-32.
- Marshall, I. N. 1989. Consciousness and Bose-Einstein condensates. *New Ideas in Psychology* 7:73-83.
- Marshall, I. N. 1995. Some phenomenological implications of a quantum model of consciousness. *Minds and Machines* 5:609-20.
- Penrose, R. 1994. Mechanisms, microtubules, and the mind. *Journal of Consciousness Studies* 1:241-49.
- Scott, A. 1996. On quantum theories of the mind. *Journal of Consciousness Studies* 3:484-91.
- Stapp, H. P. 1985. Consciousness and values in the quantum universe. *Foundations of Physics* 15:35-47.

- Stapp, H. P. 1994. Theoretical model of a purported empirical violation of the predictions of quantum mechanics. *Physical Review A* 50:18-22.
- Stapp, H. P. 1995. The hard problem: A quantum approach. *Journal of Consciousness Studies* 3:194-210. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.
- Stapp, H. P. 1997. Science of consciousness and the hard problem. *Journal of Mind and Behavior* 18:171-93.
- Triffet, T. & Green, H. S. 1996. Consciousness: Computing the uncomputable. *Mathematical and Computational Modelling* 24:37-56.
- Wolf, F. A. 1996. On the quantum mechanics of dreams and the emergence of self-awareness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Zohar, D. 1995. A quantum-mechanical model of consciousness and the emergence of 'I'. *Minds and Machines* 5:597-607.
- Zohar, D. 1996. Consciousness and Bose-Einstein condensates. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

6.3c

Consciousness and Physics, Misc

- Bilodeau, D. 1996. Physics, machines, and the hard problem. *Journal of Consciousness Studies* 3:386-401. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.
- Bohm, D. J. 1986. A new theory of the relationship of mind and matter. *Journal of the American Society for Psychical Research* 80:113-35.
- Burns, J. 1990. Contemporary models of consciousness, parts I & II. *Journal of Mind and Behavior* 11:153-171 & 12:407-420.
- Clarke, C. J. S. 1995. The nonlocality of mind. *Journal of Consciousness Studies* 2:231-40. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.
- Culbertson, J. 1982. *Consciousness: Natural and Artificial*. Libra.
- de Silva, F. 1996. Consciousness and special relativity. *IEEE Engineering in Medicine and Biology*

Magazine 15:21-26.

Dyer, M. G. 1994. Quantum physics and consciousness, creativity, and computers. *Journal of Mind and Behavior* 15:265-90.

Elitzur, A. 1996. Time and consciousness: The uneasy bearing of relativity on the mind-body problem. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Herbert, N. 1993. *Elemental Mind: Human Consciousness and the New Physics*. Dutton.

Ho, M. W. 1997. Quantum coherence and conscious experience. *Kybernetes* 26:265-76.

Hodgson, D. 1988. *The Mind Matters: Consciousness and Choice in a Quantum World*. Oxford: Oxford University Press.

Hodgson, D. 1996. Nonlocality, local indeterminism, and consciousness. *Ratio* 9:1-22.

Nair, R. 1991. Quantum physics and the philosophy of mind: An essay review. *Journal of Scientific and Industrial Research* 50:66975.

Nunn, C. M. H. , Clarke, C. J. S. & Blott, B. H. 1994. Collapse of a quantum field may affect brain function. *Journal of Consciousness Studies* 1:127-39.

Nunn, C. M. H. 1996. On the geometry of consciousness. *Journal of Consciousness Studies* 3:477-83.

Penrose, R. 1989. *The Emperor's New Mind*. Oxford University Press.

Penrose, R. 1994. *Shadows of the Mind*. Oxford University Press.

Penrose, R. 1997. *The Large, the Small, and the Human Mind*. Cambridge University Press.

Squires, E. 1990. *Conscious Mind in the Physical World*. Adam Hilger.

Zohar, D. & Marshall, I. 1990. *The Quantum Self*. Morrow.

6.4

Consciousness and Science, Misc

6.4a

Evolution of Consciousness

Arhem, P. & Liljenstrom, H. 1997. On the coevolution of consciousness and cognition. *Journal of Theoretical Biology* 187:601-12.

Barlow, H. B. 1980. Nature's joke: A conjecture on the biological role of consciousness. In (B. Josephson & V. Ramachandran, eds) *Consciousness and the Physical World*. Pergamon Press.

Barlow, H. B. 1987. The biological role of consciousness. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Cairns-Smith, A. G. 1996. *Evolving the Mind: On the Nature of Matter and the Origin of Consciousness*. Cambridge University Press.

Coan, R. W. 1989. Alternative views on the evolution of consciousness. *Journal of Human Psychology* 29:167-99.

Cotterill, R. 2000. Did consciousness evolve from self-paced probing of the environment, and not from reflexes? *Brain and Mind* 1:283-298.

Crook, J. H. 1980. *The Evolution of Human Consciousness*. Oxford University Press.

Donald, M. 1995. The neurobiology of human consciousness: An evolutionary approach. *Neuropsychologia* 33:1087-1102.

Eccles, J. C. 1992. Evolution of consciousness. *Proceedings of the National Academy of Sciences USA* 89:7320-24.

Calvin, W. H. 1991. *The Ascent of Mind: Ice Age Climates and the Evolution of Intelligence*. Bantam Books.

Crook, J. H. 1980. *The Evolution of Human Consciousness*. Oxford University Press.

Dennett, D. C. 1986. Julian Jaynes' software archaeology. *Canadian Psychology* 27:149-54.

Dewart, L. 1989. *Evolution and Consciousness: The Role of Speech in the Origin and Development of Human Nature*. University of Toronto Press.

Glynn, I. M. 1993. The evolution of consciousness: William James' unresolved problem. *Biological Reviews of the Cambridge Philosophical Society* 68:599-616.

Hameroff, S. R. 1998. Did consciousness cause the Cambrian evolutionary explosion? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Humphrey, N. 1992. *A History of the Mind: Evolution and the Birth of Consciousness*. Simon and Schuster.

Jaynes, J. 1976. *The Origins of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin.

Jonker A. 1987. The origin of the human mind. A speculation on the emergence of language and human consciousness. *Acta Biotheoretica* 36:129-77.

King, J. E. , Rumbaugh, D. M. & Savage-Rumbaugh, E. S. 1998. Evolution of intelligence, language, and other emergent processes for consciousness: A comparative perspective. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Lindahl, B. I. B. 1997. Consciousness and biological evolution. *Journal of Theoretical Biology* 187:613-29.

Nichols, S. & Grantham, T. 2000. Adaptive complexity and phenomenal consciousness. *Philosophy Of Science* 67:648-670.

Ornstein, R. 1991. *The Evolution of Consciousness: Of Darwin, Freud, and Cranial Fire: The Origins of the Way We Think*. Prentice-Hall.

Povinelli, D. J. 1987. Monkeys, apes, mirrors, minds: The evolution of self-awareness in primates. *Human Evolution* 2:493-507.

Pribram, K. H. , Jerison, H. J. , McGuinness, D. , & Eccles, J. C. 1982. The evolution of consciousness: A symposium. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Reber, A. S. 1992. The cognitive unconscious: An evolutionary perspective. *Consciousness and Cognition* 1:93-133.

Reber, A. S. 1992. An evolutionary context for the cognitive unconscious. *Philosophical Psychology* 5:33-51.

Reber, A. & Allen, R. 2000. Individual differences in implicit learning: Implications for the evolution of consciousness. In (R. Kunzendorf & B. Wallace, eds) *Individual differences in conscious experience*. John Benjamins.

Rogers, L. J. 1995. Evolution and development of brain asymmetry, and its relevance to language, tool

use and consciousness. *International Journal of Comparative Psychology* 8:1-15.

Roth, G. 2000. The evolution and ontogeny of consciousness. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Towers, B. 1979. Consciousness and the brain: Evolutionary aspects. In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.

Vandervert, L. R. 1995. Chaos theory and the evolution of consciousness and mind: A thermodynamic/holographic resolution to the mind-body problem. *New Ideas in Psychology* 13:107-27.

6.4b

Consciousness and Language

Arbib, M. A. 1972. Consciousness: The secondary role of language. *Journal of Philosophy* 69.

Bailey, W. 1986. Consciousness and action/motion theories of communication. *Western Journal of Speech Communication* 50:74-86.

Blachowicz, J. 1997. The dialogue of the soul with itself. *Journal of Consciousness Studies* 4:485-508.

Carruthers, P. 1996. The involvement of language in conscious thinking. In *Language, Thought, and Consciousness*. Cambridge University Press.

Chafe, W. L. 1980. The deployment of consciousness in the construction of narrative. In (W. Chafe, ed) *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex.

Chafe, W. L. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.

Chafe, W. L. 1996. How consciousness shapes language. *Pragmatics and Cognition* 4:35-54.

Chafe, W. 2000. A linguist's perspective on William James and "The Stream of Thought." *Consciousness & Cognition* 9:618-628.

Chapman, S. B. & Ulatowska, H. K. 1997. Discourse in dementia: Considerations of Consciousness. In (M. Stamenov, ed) *Language Structure, Discourse, and the Access to Consciousness*. John Benjamins.

de Beaugrande, R. 1997. The "conscious and unconscious mind" in the theoretical discourse of modern linguistics. In (M. Stamenov, ed) *Language Structure, Discourse, and the Access to Consciousness*. John

Benjamins.

Fludernik, M. , & Sell, R. D. 1995. The fictions of language and the languages of fiction: The linguistic representation of speech and consciousness. *Journal of Pragmatics* 24:557.

Johnston, P. K. 1997. Battle within: Shakespeare's brain and the nature of human consciousness. *Journal of Consciousness Studies* 4:365-73.

Langacker, R. W. 1997. Consciousness, construal, and subjectivity. In (M. Stamenov, ed) *Language Structure, Discourse, and the Access to Consciousness*. John Benjamins.

Lecours, A. R. 1998. Language contrivance on consciousness (and vice versa). In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Macphail, E. 2000. The search for a mental Rubicon. In (C. Heyes & L. Huber, eds) *The Evolution of Cognition*. MIT Press.

Markey, J. F. 1925. The place of language habits in a behavioristic explanation of consciousness. *Psychological Review* 32:384-401.

Pronko, N. H. 1987. Language with or without consciousness. In (G. Greenberg & E. Tobach, eds) *Cognition, Language and Consciousness: Integrative Levels*. Lawrence Erlbaum.

Ricciardelli, L. A. 1993. Two components of metalinguistic awareness: Control of linguistic processing and analysis of linguistic knowledge. *Applied Psycholinguistics* 14:349-367.

Schooler, J. W & Fiore, S. M. 1997. Consciousness and the limits of language: You can't always say what you think or think what you say. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Sekhar, A. C. 1948. Language and consciousness. *Indian Journal of Psychology* 23:79-84.

Sinha, V. 1987. Symbolic language not a pre-requisite for self-awareness. *Psycho-Lingua* 17:115-121.

Stamenov, M. I. (ed) 1997. *Language Structure, Discourse, and the Access to Consciousness*. John Benjamins.

Stamenov, M. I. 1997. Grammar, meaning, and consciousness: What sentence structure can tell us about the structure of consciousness. In (M. Stamenov, ed) *Language Structure, Discourse, and the Access to Consciousness*. John Benjamins.

6.4c

Animal Consciousness

Allen, G. E. 1987. Materialism and reductionism in the study of animal consciousness. In (G. Greenberg, E. Tobach, eds) *Cognition, Language, and Consciousness: Integrative Levels*. Lawrence Erlbaum.

Bekoff, M. 1992. Scientific ideology, animal consciousness, and animal protection: A principled plea for unabashed common sense. *New Ideas in Psychology* 10:79-94.

Bradshaw, R. H. 1998. Consciousness in nonhuman animals: Adopting the precautionary principle. *Journal of Consciousness Studies* 5:108-14.

Burghardt, G. 1985. Animal awareness: Current perceptions and historical perspective. *American Psychologist* 40:905-919.

Carruthers, P. 1989. Brute experience. *Journal of Philosophy* 258-69.

Cheney, D. L. & Seyfarth, R. M. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. University of Chicago Press.

Crook, J. H. 1983. On attributing consciousness to animals. *Nature* 303:11-14.

Dawkins, M. S. 1993. *Through Our Eyes Only: The Search for Animal Consciousness*.

Dennett, D. C. 1995. Animal consciousness: What matters and why? *Social Research* 62:691-710.

Eccles J. C. 1982. Animal consciousness and human self-consciousness. *Experientia* 38:1384-91.

Gallup, G. G. 1985. Do minds exist in species other than our own? *Neuroscience and Biobehavioral Reviews* 9:631-41.

Griffin, D. R. 1981. *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience*. William Kaufmann.

Griffin, D. R. 1985. Animal consciousness. *Neuroscience and Biobehavioral Reviews* 9:615-22.

Griffin, D. R. 1992. *Animal Minds*. University of Chicago Press.

Griffin, D. R. 1995. Windows on animal minds. *Consciousness and Cognition* 4:194-204.

- Heyes, C. M. 1987. Cognisance of consciousness in the study of animal knowledge. In (W. Callebaut & R. Pinxten, eds) *Evolutionary Epistemology: A Multiparadigm Program*. Reidel.
- Hughes, H. 2001. *Sensory Exotica: A World Beyond Human Experience*. MIT Press.
- Jolley, N. 1995. Sensation, intentionality, and animal consciousness. *Ratio* 8:128-42.
- Jolly, A. 1991. Conscious chimpanzees? A review of recent literature. In (C. Ristau, ed) *Cognitive Ethology*. Lawrence Erlbaum.
- Latto, R. 1986. The question of animal consciousness. *Psychological Record* 36:309-14.
- Oakley, D. A. 1985. Animal awareness, consciousness, and self-image. In (D. Oakley, ed) *Brain and Mind*. Methuen.
- Radner, D. & Radner, M. 1996. *Animal Consciousness*. Prometheus Books.
- Reiss, D. 1998. Cognition and communication in dolphins: A question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Ristau, C. A. 1983. Language, cognition, and awareness in animals? *Annals of the New York Academy of Sciences* 406:170-86.
- Roberts, H. 1968. Consciousness in animals and automata. *Psychological Reports* 22:1226-28.
- Rollin, B. E. 1986. Animal consciousness and scientific change. *New Ideas in Psychology* 4:141-52.
- Rollin, B. E. 1989. *The Unheeded Cry: Animal Consciousness, Animal Pain, and Science*. Oxford University Press.
- Rothschild, M. 1993. Thinking about animal consciousness. *Journal of Natural History* 27:509-12.
- Rushen, J. P. 1985. The scientific status of animal consciousness. *Applied Animal Behaviour Science* 13:387-390.
- Savage-Rumbaugh, E. S. & Rumbaugh, D. 1998. Perspectives on consciousness, language, and other emergent processes in apes and humans. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Tye, M. 1997. The problem of simple minds: Is there anything it's like to be a honeybee? *Philosophical Studies* 88:289-317.

van Rooijen, J. 1981. Are feelings adaptations? The basis of modern applied animal ethology. *Applied Animal Ethology* 7:187-89.

Weiskrantz, L. 1995. The problem of animal consciousness in relation to neuropsychology. *Behavioral Brain Research* 71:171-75.

6.4d

Animal Self-Consciousness

Byrne, R. W. & Whiten, A. 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford University Press.

Epstein, R. , Lanza, R. P. & Skinner, B. F. 1981. "Self-awareness" in the pigeon. *Science* 212:695-96.

Gallup, G. G. 1970. Chimpanzees: Self-recognition. *Science* 167:86-87.

Gallup, G. G. 1975. Toward an operational definition of self-awareness. In (R. Tuttle, ed) *Socioecology and the Psychology of Primates*. Mouton.

Gallup, G. G. 1977. Self-recognition in primates: A comparative approach to the bidirectional properties of consciousness. *American Psychologist* 32:329-38.

Gallup, G. G. 1979. Self-recognition in chimpanzees and man: A developmental and comparative perspective. In (M. Lewis & M. Rosenblum, eds) *Genesis of Behavior, Volume 2*. Plenum Press.

Gallup, G. G. 1982. Self-awareness and the emergence of mind in primates. *American Journal of Primatology* 2:237-48.

Gallup, G. G. 1987. Self-awareness. In (G. Mitchell, ed) *Comparative Primate Biology, Volume 2*. Liss.

Gallup, G. G. 1991. Toward a comparative psychology of self-awareness: Species limitations and cognitive consequences. In (G. Goethals & J. Strauss, eds) *The Self: An Interdisciplinary Perspective*. Springer-Verlag.

Gallup, G. G. 1994. Self-recognition: Research strategies and experimental design. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Hart, D. & Karmel, M. P. 1996. Self-awareness and self-knowledge in humans, apes, and monkeys. In (A.

Russon, K. Bard, & S. Parkers, eds) *Reaching into Thought: The Minds of the Great Apes*. Cambridge University Press.

Heyes, C. M. 1994. Reflections on self-recognition in primates. *Animal Behaviour* 47:909-19.

Hyatt, C. W. & Hopkins, W. 1994. Self-awareness in bonobos and chimpanzees: A comparative perspective. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Marten, K. & Psarakos, S. 1992. Using self-view television to distinguish between self-examination and social behavior in the bottlenose dolphin. *Consciousness and Cognition* 4:205-24.

Marten, K. & Psarakos, S. 1994. Evidence for self-awareness in the bottlenose dolphin. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Miles, H. L. 1994. Me Chantek: The development of self-awareness in a signing orangutan. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Moynihan, M. H. 1997. Self-awareness, with specific references to Coleoid cephalopods. In (R. Mitchell, N. Thompson, & H. Miles, eds) *Anthropomorphism, Anecdotes, and Animals*. SUNY Press.

Parker, S. T. 1991. A developmental approach to the origins of self-recognition in great apes. *Human Evolution* 6:435-49.

Patterson, F. G. P. & Cohn, R. 1994. Self-recognition and self-awareness in lowland gorillas. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Povinelli, D. J. 1987. Monkeys, apes, mirrors, minds: The evolution of self-awareness in primates. *Human Evolution* 2:493-507.

Suarez, S. D. & Gallup, G. G. 1981. Self-recognition in chimpanzee and orangutans, but not gorillas. *Journal of Human Evolution* 10:175-88.

Swartz, K. B. & Evans, S. 1991. Not all chimpanzees show self-recognition. *Primates* 32:483-96.

6.4e

Altered States of Consciousness

- Atkinson, R. P. & H. Earl. 1996. Enhanced vigilance in guided meditation: Implications of altered consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Austin, J. H. 1998. *Zen and the Brain: Toward an Understanding of Meditation and Consciousness*. MIT Press.
- Forman, R. (ed) 1990. *The Problem of Pure Consciousness: Mysticism and Philosophy*. Oxford University Press.
- Forman, R. 1998. What does mysticism have to teach us about consciousness? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness 1996*. MIT Press.
- Goleman, D. 1976. Meditation and consciousness: An Asian approach to mental health. *American Journal of Psychotherapy* 30:41-54.
- Hilgard, E. R. 1979. Consciousness and control: Lessons from hypnosis. *Australian Journal of Clinical & Experimental Hypnosis* 7:103-15.
- Hunt H. T. 1985. Cognition and states of consciousness: the necessity for empirical study of ordinary and nonordinary consciousness for contemporary cognitive psychology. *Perceptual and Motor Skills* 60:239-82.
- Katz, J. M. 1983. Altered states of consciousness and emotion. *Imagination, Cognition and Personality* 2:37-50.
- Keen, E. 2000. *Chemicals for the Mind: Psychopharmacology and Human Consciousness*. Praeger.
- Kunzendorf, R. G. , Beltz, S. M. & Tymowicz, G. 1992. Self-awareness in autistic subjects and deeply hypnotized subjects: Dissociation of self-concept versus self-consciousness. *Imagination, Cognition and Personality* 11:129-41.
- Munglani, R. & Jones, J. G. 1992. Sleep and general anesthesia as altered states of consciousness. *Journal of Psychopharmacology* 6:399-409.
- Novak, P. 1996. Buddhist meditation and consciousness of time. *Journal of Consciousness Studies* 3:267-77.
- Oakley, D. 1999. Hypnosis and consciousness: A structural model. *Contemporary Hypnosis* 16:215-223.
- Pekala, R. J. & Kumar, V. K. 1989. Phenomenological patterns of consciousness during hypnosis:

Relevance to cognition and individual differences. *Australian Journal of Clinical and Experimental Hypnosis* 17:1-20.

Pekala, R. J. & Cardena, E. 2000. Methodological issues in the study of altered states of consciousness and anomalous experiences. In (E. Cardena & S. Lynn, eds) *Varieties of Anomalous Experience: Examining the Scientific Evidence*. American Psychological Association.

Pekala, R. & Kumar, V. 2000. Individual differences in patterns of hypnotic experience across low and high hypnotically susceptible individuals. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Shapiro, D. H. 1982. Meditation as an altered state of consciousness: Contributions of Western behavioral science. *Journal of Transpersonal Psychology* 15:61-81.

Spivak, L. , V. Puzenko, S. Medvedev, & Y. Polyakov 1990. Neurophysiological correlates of the altered state of consciousness during hypnosis. *Human Physiology* 16:405-410.

Tart, C. T. (ed) 1990. *Altered States of Consciousness* (third edition). Harper Collins.

Tart, C. T. 1998. Transpersonal psychology and methodologies for a comprehensive science of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Tart, C. T. 2000. Investigating altered states of consciousness on their own terms: State-specific sciences. In (M. Velmans, eds) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

Tinnin, L. 1990. Mental unity, altered states of consciousness, and dissociation. *Dissociation: Progress in the Dissociative Disorders* 3:154-59.

Travis, F. & Pearson, C. 2000. Pure consciousness: Distinct phenomenological and physiological correlates of "consciousness itself". *International Journal of Neuroscience* 100:77-89.

West, M. 1983. Meditation and self-awareness: Physiological and phenomenological approaches. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Wolman, B. B. & Ullman, U. 1986. *Handbook of States of Consciousness*. van Nostrand Reinhold.

Venkatesh S. , Raju T. R. , Shivani, Y. , Tompkins G. , & Meti B. L. 1997. A study of structure of phenomenology of consciousness in meditative and non-meditative states. *Indian Journal of Physiology and Pharmacology* 41:149-53.

Walsh, R. 1998. States and stages of consciousness: Current research and understanding. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

6.4f

Parapsychology and Consciousness

Beloff, J. 1973. The subliminal and the extrasensory. *Parapsychology Review* 4:23-27.

Beloff, J. 1976. Mind-body interactionism in light of the parapsychological evidence. *Theoria to Theory* 10:125-37.

Beloff, J. 1980. Could there be a physical explanation for psi? *Journal of the Society for Psychical Research* 50:263-272

Beloff, J. 1987. Parapsychology and the mind-body problem. *Inquiry* 30:215-25.

Beloff, J. 1989. Dualism: A parapsychological perspective. In (J. Smythies & J. Beloff, eds) *The Case for Dualism*. Virginia University Press.

Bem, D. J. & Honorton, C. 1994. Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin* 115:4-18.

Bierman, D. 1998. Do psi phenomena suggest radical dualism? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Blackmore, S. 1991. Psi in science. *Journal of the Society for Psychical Research* 57:404-11.

Blackmore, S. 1998. Why psi tells us nothing about consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Bohm, D. J. 1986. A new theory of the relationship of mind and matter. *Journal of the American Society for Psychical Research* 80:113-35.

Braud, W. G. 1994. The role of mind in the physical world: A psychologist's view. *European Journal of Parapsychology* 10:66-77.

Braude, S. E. 1979. *ESP and Psychokineses: A Philosophical Examination*. Temple University Press.

Braude, S. E. 1986. *The Limits of Influence: Psychokinesis and the Philosophy of Science*. Routledge and Kegan Paul.

- Burns, J. E. 1993. Current hypotheses about the nature of the mind-brain relationship and their relationship to findings in parapsychology. In (K. Rao, ed) *Cultivating Consciousness*. Praeger.
- Burns, J. E. 1993. Time, consciousness, and psi. In (B. Kane, J. Millay, & D. H. Brown, eds) *Silver Threads: 25 Years of Parapsychology Research*. Praeger.
- Burns, J. E. 1986. Consciousness and psi. *PSI Research* 5:166-205.
- Dilley, F. B. 1989. Mind-brain interaction and psi. *Southern Journal of Philosophy* 26:469-80.
- Dilley, F. B. 1990. Telepathy and mind-brain dualism. *Journal of the Society for Psychical Research* 56:129-37.
- Edge, H. L. 1989. Psi, self, and the new mentalism. In (L. Henkel & J. Palmer, eds) *Research in Parapsychology 1989*. Scarecrow Press.
- Eisenbud, J. 1975. The mind-matter interface. *Journal of the American Society for Psychical Research* 69:115-26.
- Goswami, A. 1986. The quantum theory of consciousness and psi. *PSI Research* 5:145-65.
- Griffin, D. R. 1993. Parapsychology and philosophy: A Whiteheadian postmodern perspective. *Journal of the American Society for Psychical Research* 87:217-88.
- Griffin, D. R. 1994. Dualism, materialism, idealism, and psi: A reply to John Palmer. *Journal of the American Society of Psychical Research* 88:23-39.
- Grof, S. 2000. *Psychology of the Future: Lessons from Modern Consciousness Research*. State University of New York Press.
- Heath, P. 2000. The PK zone: A phenomenological study. *Journal of Parapsychology* 64:53-72.
- Honorton, C. 1985. Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology* 1:51-91.
- Hubbard, T. L. 1996. Consciousness and cognition beyond the body: Functionalist cognitive science and the possibility of out-of-body experiences and reincarnation. *Journal of the American Society for Psychical Research* 90:202-20.
- Hyman, R. 1985. The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology* 49:3-49.

- Hyman, Ray & Honorton, C. 1986. A joint communique: The psi ganzfeld controversy. *Journal of Parapsychology* 50:351-64.
- Hyman, R. 1994. Anomaly or artifact? Comments on Bem and Honorto. *Psychological Bulletin* 115:19-24.
- Jahn, R. G. & Dunne, B. J. 1987. *Margins of Reality: The Role of Consciousness in the Physical World*. Harcourt Brace Jovanovich.
- Kreitler, H. & Kreitler, S. 1973. Subliminal perception and extrasensory perception. *Journal of Parapsychology* 37:163-88.
- Krippner, S. & George, L. 1986. Psi phenomena as related to altered states of consciousness. In (B. Wolman & M. Ullman, eds) *Handbook of States of Consciousness*. van Nostrand Reinhold.
- Mattuck, R. 1982. A crude model of the mind-matter interaction using Bohm-Bub hidden variables. *Journal of the Society for Psychical Research* 51:238-245.
- Nash, C. B. 1976. Psi and the mind-body problem. *Journal of the Society for Psychical Research* 48:267-70.
- Nash, C. B. 1995. A panpsychic theory of mind and matter. *Journal of the Society for Psychical Research* 60:171-73.
- Poynton, J. C. 1994. Making sense of psi: Whitehead's multilevel ontology. *Journal of the Society for Psychical Research* 59:401-12.
- Price, E. A. 1981. A "three worlds" perspective to the mind-brain relationship in parapsychology. *Parapsychological Journal of South Africa* 2:38-49.
- Rao, K. R. & Palmer, J. 1987. The anomaly called psi: Recent research and criticism. *Behavioral and Brain Sciences* 10:539-51.
- Rao, K. R. 1991. Consciousness research and psi. *Journal of Parapsychology* 55:1-43.
- Rauscher, E. A. 1983. Multidimensional properties of consciousness and some laws of reality. *PSI Research* 2:53-66.
- Richards, D. G. 1996. Psi and the spectrum of consciousness. *Journal of the American Society for Psychical Research* 90:251-67.

- Roberts, F. S. 1991. Some apparently non-cerebral aspects of consciousness. *Journal of the Society for Psychical Research* 58:31-38.
- Roberts, F. S. 1995. Is physically-based consciousness a reality? *Journal of the Society for Psychical Research* 60:398-400.
- Roney-Dougal, S. M. 1986. Subliminal and psi perception: A review of the literature. *Journal of the Society for Psychical Research* 53:405-34.
- Smythies, J. M. 1960. Three classical theories of mind. *Journal of the Society for Psychical Research* 40:385-397.
- Stokes, D. M. 1982. On the relationship between mind and brain. *Parapsychology Review*, 13:22-27.
- Stokes, D. M. 1993. Mind, matter, and death: Cognitive neuroscience and the problem of survival. *Journal of the American Society for Psychical Research* 87:41-84.
- Stokes, D. M. 1997. *The Nature of Mind: Parapsychology and the Role of Consciousness in the Physical World*. McFarland and Co.
- Tiller, W. , Kohane, M. , & Dibble, W. 2000. Can an aspect of consciousness be imprinted into an electronic device? *Integrative Physiological & Behavioral Science* 35:142-163.
- Varvoglis, M. 1996. Nonlocality on a human scale: Psi and consciousness research. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Wade, J. 1998. Physically transcendent awareness: A comparison of the phenomenology of consciousness before birth and after death. *Journal of Near-Death Studies* 16:249-275.
- Walker, E. H. 1984. A review of criticisms of the quantum-mechanical theory of psi phenomena. *Journal of Parapsychology* 48:277-32.
- Woodworth, H. 1942. Report of investigations into an obscure function of the subconscious mind. *Journal of the American Society for Psychical Research* 36:185-230.

6.4g

Phenomenology

Ackerman, D. 1990. *A Natural History of the Senses*. Randhom House,

- Arvidson, P. S. 1992. On the origin of organization in consciousness. *Journal of the British Society of Phenomenology* 23:53-65.
- Arvidson, P. S. 1996. Toward a phenomenology of attention. *Human Studies* 19:71-84.
- Baars, B. J. 1993. Putting the focus on the fringe: Three empirical cases. *Journal of Consciousness Studies* 2:126-36.
- Chokr, N. N. 1992. Mind, consciousness, and cognition: Phenomenology vs. cognitive science. *Husserl Studies* 9:179-97.
- Deikman, A. 1996. 'I' = awareness. *Journal of Consciousness Studies* 3:350-56.
- Depraz, N. , Varela, F. , & Vermersch, P. 2000. The gesture of awareness: An account of its structural dynamics. In (M. Velmans, ed) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.
- de Quincey, C. 2000. Intersubjectivity: Exploring consciousness from the second-person perspective. *Journal of Transpersonal Psychology* 32:135-155.
- Diaz, J. 1996. The stream revisited: A process model of phenomenological consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Ellis, R. 1986. *An Ontology of Consciousness*. Kluwer.
- Galin, D. 1994. The structure of awareness: Contemporary applications of William James' forgotten concept of "the fringe". *Journal of Mind and Behavior* 15:375-401.
- Galin, D. 1996. The structure of subjective experience: Sharpen the concepts and terminology. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Gallagher, S. 1997. Mutual enlightenment: Recent phenomenology in cognitive science. *Journal of Consciousness Studies* 4:195-214.
- Gurwitsch, A. 1964. *The Field of Consciousness*. Duquesne University Press.
- Gurwitsch, A. 1966. *Studies in Phenomenology and Psychology*. Northwestern University Press.
- Ihde, D. 1977. *Experimental Phenomenology*. Putnam.
- Jopling, D. A. 1996. Sub-phenomenology. *Human Studies* 19:153-73.

- Koestenbaum, P. 1962. The sense of subjectivity. *Review of Existential Psychology* 2:47-65.
- Lind, R. 1996. Micro-phenomenology: Toward a hypothetico-inductive science of experience. *International Philosophical Quarterly* 36:429-42.
- Mangan, B. 1993. Taking phenomenology seriously: The "fringe" and its implication for cognitive research. *Consciousness and Cognition* 2:89-108.
- Marbach, E. 1993. *Mental Representation and Consciousness: Toward a Phenomenological Theory of Representation and Reference*. Kluwer.
- Marbach, E. 1996. Understanding the representational mind: A phenomenological perspective. *Human Studies* 19:137-52.
- Marbach, E. 2000. The place for an Ego in current research. In (D. Zahavi, ed) *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience*. John Benjamins.
- Natsoulas, T. 1997. The presence of environmental objects to perceptual consciousness: An integrative, ecological and phenomenological approach. *Journal of Mind & Behavior* 18:371-390.
- Natsoulas, T. 1997. The presence of environmental objects to perceptual consciousness: A difference it makes for psychological functioning. *American Journal of Psychology* 110:507-526.
- Nelson, P. 1998. Consciousness as reflexive shadow: An operational psychophenomenological model. *Imagination, Cognition and Personality* 17:215-228.
- Pekala, R. J. , Wenger C. F. , & Levine R. L. 1985. Individual differences in phenomenological experience: states of consciousness as a function of absorption. *Journal of Personality and Social Psychology* 48:125-32.
- Pekala, R. J. & Levine, R. L. 1982. Mapping consciousness: Development of an empirical-phenomenological approach. *Imagination, Cognition & Personality* 1:29-47.
- Rao, K. R. 1998. Two faces of consciousness: A look at Eastern and Western perspectives. *Journal of Consciousness Studies* 5:309-27.
- Shanon, B. 1984. The case for introspection. *Cognition and Brain Theory* 7:167-80.
- Shear, J. 1996. The hard problem: Closing the empirical gap. *Journal of Consciousness Studies* 3:54-68. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press,

Stevens, R. 2000. Phenomenological approaches to the study of conscious awareness. In (M. Velmans, ed) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

Wilber, K. 2000. Waves, streams, states and self: Further considerations for an integral theory of consciousness. *Journal of Consciousness Studies* 7:145-176.

Varela, F. 1995. Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies* 3:330-49. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.

6.4h

Foundations

Baars, B. J. 1994. A thoroughly empirical approach to consciousness. *Psyche* 1.

Baruss, I. , & Moore, R. J. 1992. Measurement of beliefs about consciousness and reality. *Psychological Reports* 71:59-64.

Battista, J. R. 1978. The science of consciousness. In (K. S. Pope & J. L. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Block, N. 2001. Paradox and cross purposes in recent work on consciousness. *Cognition* 79:197-219.

Conrad, D. 1996. Consciousness, privacy, and information. *Biosystems* 38:207-10.

Dennett, D. 2001. Are we explaining consciousness yet? *Cognition* 79:221-37.

Dunlop, K. 1912. The case against introspection. *Psychological Review* 19:404-13.

Flanagan, O. J. 1995. Consciousness and the natural method. *Neuropsychologia* 33:1103-15.

Foss, J. 2000. *Science and the Riddle of Consciousness: A Solution*. Kluwer Academic Publishers.

Goldman, A. 1997. Science, publicity, and consciousness. *Philosophy of Science* 64:525-45.

Goldman, A. 2000. Can science know when you're conscious? Epistemological foundations of consciousness research. *Journal Of Consciousness Studies* 7:3-22.

Grinker, R. R. 1953. Problems of consciousness: A review, an analysis, and a proposition. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Fourth Conference*. Josiah Macy

Foundation.

Jack, A. I. & Shallice, T. 2001. Introspective physicalism as an approach to the science of consciousness. *Cognition* 79:161-196.

Lyons, W. 1986. *The Disappearance of Introspection*. MIT Press.

Miller, D. 2000. Designing a bridge for consciousness: Are criteria for a unification of approaches feasible? *Advances in Mind-Body Medicine* 16:82-89.

Nunez, R. 1997. Eating soup with chopsticks: Dogmas, difficulties, and alternatives in the study of conscious experience. *Journal of Consciousness Studies* 4:143-66.

Pekala, R. & Cardena, E. 2000. Methodological issues in the study of altered states of consciousness and anomalous experiences. In (E. Cardena & S. Lynn, eds) *Varieties of Anomalous Experience: Examining the Scientific Evidence*. American Psychological Association.

Scott, A. C. 1998. Reductionism revisited. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Scott, A. 2000. Modern science and the mind. In (M. Velmans, ed) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

Searle, J. 1998. How to study consciousness scientifically. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Simon, H. A. 1997. Scientific approaches to the question of consciousness. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Stevens, S. S. 1966. Quantifying the sensory experience. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Varela, F. 1998. A science of consciousness as if experience mattered. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness 1996*. MIT Press.

Velmans, M. 1994. A reflexive science of consciousness. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

Velmans, M. 1996. Introduction to the science of consciousness. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Velmans, M. 1998. Goodbye to reductionism: Complementary first and third-person approaches to consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Velmans, M. 2000. *Understanding Consciousness*. Routledge.

Wallace, B. 2000. *The Taboo of Subjectivity: Toward a New Science of Consciousness*. Oxford University Press.

Williams, D. C. 1934. Scientific method and the existence of consciousness. *Psychological Review* 41:461-79.

6.4i

Consciousness and Science, Misc

Abramson, H. A. (ed) 1950. *Problems of Consciousness: Transactions of the First Conference*. Josiah Macy Foundation.

Abramson, H. A. (ed) 1951. *Problems of Consciousness: Transactions of the Second Conference*. Josiah Macy Foundation.

Abramson, H. A. (ed) 1952. *Problems of Consciousness: Transactions of the Third Conference*. Josiah Macy Foundation.

Abramson, H. A. (ed) 1953. *Problems of Consciousness: Transactions of the Fourth Conference*. Josiah Macy Foundation.

Abramson, H. A. (ed) 1954. *Problems of Consciousness: Transactions of the Fifth Conference*. Josiah Macy Foundation.

Bielecki, A. , Kokoszka, A. , & Holas, P. 2000. Dynamic systems theory approach to consciousness. *International Journal of Neuroscience* 104:29-47.

Blakemore, C. & Greenfield, S. 1987. *Mindwaves: Thoughts on Intelligence, Identity, and Consciousness*. Blackwell.

Bock, G. R. & Marsh, J. (eds) 1993. *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

Cohen, J. D. & Schooler, J. W. (eds) 1997. *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

- Cornwell, J. (ed) 1998. *Consciousness and Human Identity*. Oxford University Press.
- Cotterill, R. 2000. *Enchanted Looms: Conscious Networks in Brains and Computers*. Cambridge University Press.
- Hameroff, S. R, Kaszniak, A. & Scott, A. (eds) 1996. *Toward a Science of Consciousness: The First Tucson Discussions and Debates*. MIT Press.
- Ito, M. , Miyashita, Y. , & Rolls, E. T. (eds) 1997. *Cognition, Computation, and Consciousness*. Oxford University Press.
- Jarvilehto, T. 2000. The theory of the organism-environment system: The problem on mental activity and consciousness. *Integrative Physiological & Behavioral Science* 35:35-57.
- John, E. R. 2001. A field theory of consciousness. *Consciousness and Cognition* 10:184-213.
- Josephson, B. & Ramachandran, V. S. (eds) 1980. *Consciousness and the Physical World*. Pergamon Press.
- Keyes, C. D. 1999. *Brain Mystery Light and Dark: The Rhythm and Harmony of Consciousness*. Routledge.
- Marcel, A. J. & Bisiach, E. (eds) 1988. *Consciousness in Contemporary Science*. Oxford University Press.
- Oakley, D. A. (ed) 1985. *Brain and Mind*. Methuen.
- Scott, A. 1995. *Stairway to the Mind: The Controversial New Science of Consciousness*. Springer.
- Sugarman, A. A. & Tarter, R. E. (eds) 1978. *Expanding Dimensions of Consciousness*. Springer.
- Torey, Z. 1999. *The Crucible of Consciousness*. Oxford University Press.
- Velmans, M. (eds) 1996. *The Science of Consciousness: Tutorial Essays*. Routledge.
- Wilber, K. 2000. *Integral Psychology: Consciousness, Spirit, Psychology, Therapy*. Shambhala.

Part 1: Consciousness and Qualia [1082]

=====

- 1.1 Consciousness [85]
 - 1.1a General [47]
 - 1.1b The Concept of Consciousness [38]
- 1.2 Explaining Consciousness? [115]
 - 1.2a Subjectivity and Objectivity (Nagel) [46]
 - 1.2b The Explanatory Gap (Levine) [16]
 - 1.2c 'Hard' and 'Easy' Problems (Chalmers) [25]
 - 1.2d Cognitive Closure (McGinn) [16]
 - 1.2e Miscellaneous [12]
- 1.3 Materialism and Dualism [213]
 - 1.3a The Knowledge Argument (Jackson) [46]
 - 1.3b Zombies & Modal Arguments [38]
 - 1.3c Essentialist Arguments (Kripke) [31]
 - 1.3d Arguments from Disembodiment [17]
 - 1.3e Consciousness and Physicalism, Misc [14]
 - 1.3f Consciousness and Dualism [16]
 - 1.3g Mind-Body Problem, Misc [51]
- 1.4 Specific Views on Consciousness [227]
 - 1.4a Higher-Order Thought Approaches (Rosenthal, etc) [46]
 - 1.4b Dennett on Consciousness [60]
 - 1.4c Functionalism [8]
 - 1.4d Eliminativism [16]
 - 1.4e Epiphenomenalism [30]
 - 1.4f Interactionism [16]
 - 1.4g Panpsychism [24]
 - 1.4h Intrinsic Monism (Russell, Feigl, etc) [27]
- 1.5 Consciousness and Content [129]
 - 1.5a Consciousness and Intentionality (Searle, etc) [23]
 - 1.5b The Content of Experience [48]
 - 1.5c Representationalism [38]
 - 1.5d Internalism and Externalism about Experience [12]
 - 1.5e Miscellaneous [8]
- 1.6 Aspects of Consciousness [182]
 - 1.6a Self-Consciousness [54]
 - 1.6b Unity of Consciousness [45]
 - 1.6c Homogeneity of Consciousness (Sellars, etc) [11]
 - 1.6d Knowledge of Consciousness [50]
 - 1.6e The Function of Consciousness [22]
- 1.7 Qualia [131]
 - 1.7a Qualia, General [27]
 - 1.7b Qualia and Materialism [13]
 - 1.7c Eliminativism about Qualia [11]
 - 1.7d The Inverted Spectrum [31]
 - 1.7e Absent Qualia (Block, etc) [15]
 - 1.7f Introspection and Absent Qualia (Shoemaker) [10]
 - 1.7g Functionalism and Qualia, General [24]

1.1 Consciousness

1.1a Consciousness -- General

Armstrong, D.M. & Malcolm, N. 1984. _Consciousness and Causality: A Debate on the Nature of Mind_. Blackwell.

Block, N., Flanagan, O., & Guzeldere, G. (eds) 1997. _The Nature of Consciousness: Philosophical Debates_. MIT Press.

An anthology of central philosophical papers on consciousness.

Carruthers, P. 2000. Phenomenal Consciousness: A Naturalistic Theory.
Cambridge University Press.

Catalano, J. 2000. Thinking Matter: Consciousness from Aristotle to Putnam and Sartre. Routledge.

Chalmers, D.J. 1991. Consciousness and cognition. Manuscript.
Exploring the link between consciousness and judgments about consciousness.
Coherence between these => consciousness depends on the functional but isn't
reducible. Toward a dual-aspect theory based on pattern and information.

Chalmers, D.J. 1996. The Conscious Mind: In Search of a Fundamental Theory.
Oxford University Press.
Argues against the reductive explanation of consciousness, and for a kind of
naturalistic dualism. Moves toward a "fundamental theory" to bridge the
gap, and draws out some consequences.

Churchland, P.M. & Churchland, P.S. 1997. Recent work on consciousness:
Philosophical, theoretical, and empirical. *Seminars in Neurology* 17:179-86.

Davies, M. & Humphreys, G. 1993. Consciousness: Philosophical and
Psychological Essays. Blackwell.
A collection of 5 psychological and 8 philosophical essays on consciousness.

Flanagan, O.J. 1991. Consciousness. In The Science of the Mind. MIT Press.
On the mysteries of consciousness. Argues with epiphenomenalism, "conscious
inessentialism", and the "new mysterians" (Nagel, McGinn). Toward a
naturalistic theory, drawing on ideas of Edelman, Calvin, Dennett.

Flanagan, O.J. 1992. Consciousness Reconsidered. MIT Press.
Argues that consciousness can be accounted for in a naturalistic framework.
With arguments against eliminativism and epiphenomenalism, evidence from
neuroscience and psychology, and discussions of the stream and the self.

Flanagan, O.J. & Guzeldere, G. 1997. Consciousness: A philosophical tour.
In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) Cognition, Computation, and
Consciousness. Oxford University Press.

Foss, J. 2000. Science and the Riddle of Consciousness: A Solution. Kluwer.

Graham, G. & Horgan, T. 1998. Sensations and grain processes. In (G.
Mulhauser, ed) Evolving Consciousness. John Benjamins.

Gregory, R.L. 1988. Consciousness in science and philosophy: conscience and
con-science. In (A. Marcel & E. Bisiach, eds) Consciousness in Contemporary
Science. Oxford University Press.

Guzeldere, G. 1995. Consciousness: what it is, how to study it, what to learn
from its history. *Journal of Consciousness Studies* 2:30-51.
A history of the study of consciousness, especially in psychology.

Guzeldere, G. 1995. Problems of consciousness: A perspective on contemporary
issues, current debates. *Journal of Consciousness Studies* 2:112-43.
A summary of recent philosophical debates over consciousness, focusing on
the "what/where/who/why/how" questions, the explanatory gap, and the
stalemate between "essentialist" and "causal" intuitions.

Hannay, A. 1987. The claims of consciousness: A critical survey. *Inquiry*
30:395-434.

Hannay, A. 1990. Human Consciousness. Routledge.

- Hurley, S. 1998. *Consciousness in Action*. Harvard University Press.
- Jackendoff, R. 1987. *Consciousness and the Computational Mind*. MIT Press.
Separates computational mind from phenomenological mind, and studies the former, a third-person approach. The residue is the "Mind-Mind" problem. Consciousness supervenes on an intermediate level of representation. Elegant.
- Kirk, R. 1994. *Raw Feeling: A Philosophical Account of the Essence of Consciousness*. Oxford University Press.
Physicalism can explain consciousness in all its glory. Argues against zombies and inverted-spectrum scenarios, and suggests that the explanatory gap can be bridged by an account of directly-active information-processing.
- Levine, J. 1997. Recent work on consciousness. *American Philosophical Quarterly* 34:379-404.
- Lycan, W.G. 1987. *Consciousness*. MIT Press.
- Lycan, W.G. 1996. *Consciousness and Experience*. MIT Press.
- Metzinger, T. (ed) 1995. *Conscious Experience*. Ferdinand Schoningh.
An excellent collection of 20 philosophical papers on consciousness.
- Murata, J. 1997. Consciousness and the mind-body problem. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.
- Nelkin, N. 1996. *Consciousness and the Origins of Thought*. Cambridge University Press.
- O'Shaughnessy, B. 2000. *Consciousness and the World*. Oxford University Press.
- Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. MIT Press.
- Revonsuo, A. & Kamppinen, M. (eds) 1994. *Consciousness in Philosophy and Cognitive Neuroscience*. Lawrence Erlbaum.
- Sayre, K.M. 1969. *Consciousness: A Philosophic Study of Minds and Machines*. Random House.
- Seager, W.E. 1999. *Theories of Consciousness: An Introduction and Assessment*. Routledge.
- Searle, J.R. 1989. Consciousness, unconsciousness, and intentionality. *Philosophical Topics* 17:193-209.
Argues that the first-person view has been ignored too much in the philosophy of mind. Even unconscious states are only mental by virtue of their potential consciousness.
- Searle, J.R. 1992. *The Rediscovery of the Mind*. MIT Press.
On the centrality of consciousness to the mind. Consciousness is irreducible but biological. On the history of the field, the structure of consciousness, its role in constituting intentionality, and problems with computation.
- Searle, J.R. 1993. The problem of consciousness. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.
On the notion of consciousness, its relation to the brain, and some features that need to be explained: its subjectivity, unity, intentionality, center and periphery, Gestalt structure, aspect of familiarity, and so on.
- Sheets-Johnstone, M. 1998. Consciousness: A natural history. *Journal of Consciousness Studies* 5:260-94.

Siewert, C. 1998. *The Significance of Consciousness*. Princeton University Press.

Smith, D.W. 1992. Consciousness in action. *Synthese* 90:119-43.

Sprigge, T.L.S. 1982. The importance of subjectivity: An inaugural lecture. *Inquiry* 25:143-63.

Value is only found within streams of consciousness. Three ways of studying it: phenomenology, anthropology, and by relation to the physical. With an analysis of the "self-transcending" nature of conscious intentionality.

Strawson, G. 1994. *Mental Reality*. MIT Press.

Sturgeon, S. 2000. *Matters of Mind: Consciousness, Reason and Nature*. Routledge.

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press.

Tye, M. 2000. *Consciousness, Color, and Content*. MIT Press.

Velmans, M. 1996. *The Science of Consciousness: Psychological, Neuropsychological, and Clinical Reviews*. Routledge.

Verges, F.G. 1974. Jackson on incorrigibility. *Australasian Journal of Philosophy* 52:243-50.

Villaneuva, E. (ed) 1991. *Consciousness: Philosophical Issues*. Ridgeview. A collection of philosophical articles on consciousness.

Young, A.W. & Block, N. 1997. Consciousness. In (V. Bruce, ed) *Unsolved Mysteries of the Mind: Tutorial Essays in Cognition*. Taylor and Francis.

1.1b The Concept of Consciousness

Armstrong, D.M. 1979. Three types of consciousness. In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.

Armstrong, D.M. 1981. What is consciousness? In *The Nature of Mind*. Cornell University Press.

On minimal consciousness, perceptual consciousness, and introspective consciousness. Introspective consciousness seems so special because it gives inner awareness of self, and memory of other mental events.

Baruss, I. 1986. Meta-analysis of definitions of consciousness. *Imagination, Cognition, and Personality* 6:321-29.

Bisiach, E. 1988. The (haunted) brain and consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Distinguishes C1 (phenomenal experience) from C2 (access of parts of a system to other parts). C2 is can be scientifically studied, and has a graspable, if fragmented, causal role. C1 is mysterious and perhaps beyond science.

Burt, C. 1962. The concept of consciousness. *British Journal of Psychology* 53:229-42.

Cam, P. 1985. Phenomenology and speech dispositions. *Philosophical Studies* 47:357-68.

Reportability is not phenomenology, as blindsight has reportability but no phenomenology.

Chalmers, D.J. 1997. Availability: The cognitive basis of experience? In (N. Block, O. Flanagan, and G. Guzeldere, eds) *The Nature of Consciousness*. MIT Press.

Argues that the cognitive correlate of consciousness is direct availability for global control.

Church, J. 1998. Two sorts of consciousness? *Communication and Cognition* 31:51-71.

Finkelstein, D.H. 1999. On the distinction between conscious and unconscious states of mind. *American Philosophical Quarterly* 36:79-100.

Gennaro, R.J. 1995. Does mentality entail consciousness? *Philosophia* 24:331-58.

Girle, R.A. 1996. Shades of consciousness. *Minds and Machines* 6:143-57.

Kirk, R. 1992. Consciousness and concepts. *Aristotelian Society Supplement* 66:23-40.

Analyzes consciousness in terms of the "presence" of information to the main decision-making processes of a system. No great conceptual capacities required, no higher-order thoughts. With application to blindsight.

Lormand, E. 1995. What qualitative consciousness is like. Manuscript.

Lormand, E. 1996. Nonphenomenal consciousness. *Nous* 30:242-61.

Manson, N. 2000. State consciousness and creature consciousness: A real distinction. *Philosophical Psychology* 13:405-410.

Matthews, G. 1977. Consciousness and life. *Philosophy* 52:13-26.

McBride, R. 1999. Consciousness and the state/transitive/creature distinction. *Philosophical Psychology* 12:181-196.

Moody, T.C. 1986. Distinguishing consciousness. *Philosophy and Phenomenological Research* 47:289-95.

Separates consciousness from the mental -- functionalist accounts work for the latter but not the former. With remarks on Zen "pure consciousness".

Natsoulas, T. 1978. Consciousness. *American Psychologist* 33:906-14.

On the role of consciousness in psychology, and distinguishing various notions of consciousness: mutual knowledge, internal knowledge, awareness, direct awareness, personal unity, wakefulness, and double consciousness.

Natsoulas, T. 1983. A selective review of conceptions of consciousness with special reference to behavioristic contributions. *Cognition and Brain Theory* 6:417-47.

Ideas about consciousness from Locke, Brentano, Hebb, Dennett, Skinner, Sellars, Aristotle, Gibson. Theories: inner eye vs. verbal vs. outer eye.

Natsoulas, T. 1983. Concepts of consciousness. *Journal of Mind and Behavior* 4:195-232.

Natsoulas, T. 1991. The concept of consciousness(1): The interpersonal meaning. *Journal for the Theory of Social Behavior* 21:63-89.

Natsoulas, T. 1991. The concept of consciousness(2): The personal meaning. *Journal for the Theory of Social Behavior* 21:339-67.

Natsoulas, T. 1992. The concept of consciousness(3): The awareness meaning. *Journal for the Theory of Social Behavior* 2:199-25.

Natsoulas, T. 1993. Consciousness(4): Varieties of intrinsic theory. *Journal of Mind and Behavior* 14:107-32.

Natsoulas, T. 1994. The concept of consciousness(4): The reflective meaning. *Journal for the Theory of Social Behavior* 24:373-400.

Natsoulas, T. 1994. The concept of consciousness(5): The unitive meaning. *Journal for the Theory of Social Behavior* 24:401-24.

Natsoulas, T. 1995. Consciousness(3) and Gibson's concept of awareness. *Journal of Mind and Behavior* 3:305-28.

Natsoulas, T. 1996-1998. The case for intrinsic theory (parts 1-3). *Journal of Mind and Behavior* 17:267-85, 17:369-89, 19:1.

Natsoulas, T. 1997. Consciousness and self-awareness: Consciousness(1,2,3,4,5,6). *Journal of Mind and Behavior* 18:53-94.

Nelkin, N. 1987. What is it like to be a person? *Mind and Language* 21:220-41. Critiques three senses of consc: awareness, verbalization and phenomenology. Argues that none are sufficient for person-consciousness. Quite good.

Nelkin, N. 1993. What is consciousness? *Philosophy of Science* 60:419-34. On three senses of consciousness: phenomenality, intentionality, and introspectibility. Argues from empirical evidence (especially blindsight cases) that these three are all dissociable.

O'Shaughnessy, B. 1991. The anatomy of consciousness. In (E. Villanueva, ed) *_Consciousness_*. Ridgeview.

Place, U.T. 1992. Two concepts of consciousness: The biological/private and the linguistic/social. *Acta Analytica* 7:53-72.

Rosenthal, D.M. 1990. The independence of consciousness and sensory quality. In (E. Villanueva, ed) *_Consciousness_*. Ridgeview. Argues that consciousness and sensory quality are independent properties: there can be unconscious sensations. Consciousness is a relational property.

Rosenthal, D.M. 1994. State consciousness and transitive consciousness. *Consciousness and Cognition* 2:355-63.

Shanon, B. 1990. Consciousness. *Journal of Mind and Behavior* 11:137-51. On three kinds of consciousness -- sensed being, mental awareness, and reflection -- and their relationships.

Tye, M. 1996. The burning house. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh. Uses various puzzles cases to distinguish higher-order consciousness, discriminatory consciousness, responsive consciousness, and phenomenal consciousness.

1.2 Explaining Consciousness?

1.2a Subjectivity and Objectivity (Nagel) [see also 1.3a]

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 4:435-50. Reprinted in *_Mortal Questions_* (Cambridge University Press, 1979). Physicalist explanations leave out consciousness, i.e. what it is like to be an organism. Objective accounts omit points of view (could there be an objective phenomenology?). Physicalism may be true, but we can't see how.

Nagel, T. 1979. Subjective and objective. In *Mortal Questions*. Cambridge University Press.

Subjective and objective views clash e.g. on meaning of life, free will, personal identity, mind-body problem, ethics. How to reconcile: reduction, elimination, annexation? Maybe just let multiple viewpoints coexist.

Nagel, T. 1986. *The View From Nowhere*. Oxford University Press.

Seeing philosophy as a clash between the subjective and objective views of various phenomena (mental states, self, knowledge, freedom, value, ethics). Eliminating the subjective is impossible.

Akins, K. 1993. What is it like to be boring and myopic? In (B. Dahlbom, ed) *Dennett and his Critics*. Blackwell.

Gives a detailed account of perceptual processing in bats, and suggests that we can know what bat-experience is like: it's like nerd experience. But then is there an unexplained residue?

Akins, K. 1993. A bat without qualities? In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

On what science tells us about the experience of bats, birds, and others. Why a movie of bat-experience isn't good enough -- because of the inseparability of intentionality and experience. Science can do OK.

Baker, L.R. 1998. The first-person perspective: A test for naturalism. *American Philosophical Quarterly* 35:327-348.

Biro, J.I. 1991. Consciousness and subjectivity. In (E. Villanueva, ed) *Consciousness*. Ridgeview.

No real problems are posed by subjectivity and points of view, no matter how they are construed (fixed, portable, tokens, types). It's either a confusion or a triviality about the logic of indexicality.

Biro, J.I. 1993. Consciousness and objectivity. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Carruthers, P. 2000. Sympathy and subjectivity. *Australasian Journal of Philosophy* 77:465-482.

Davis, L. 1982. What is it like to be an agent? *Erkenntnis* 18:195-213.

On what is required for consciousness of agency (rather than qualia): belief, intention, and most importantly desire, enabling a capacity to care. A robot could have all this, and it would be like something to be it.

Flanagan, O.J. 1985. Consciousness, naturalism and Nagel. *Journal of Mind and Behavior* 6:373-90.

Naturalism can do autophenomenology just fine.

Foss, J.E. 1989. On the logic of what it is like to be a conscious subject. *Australasian Journal of Philosophy* 67:305-320.

A Super Neuroscientist will know how we describe and think about experience, so will know as much as a Super Sympathist. One doesn't have to imagine to know what it's like. With remarks on bat experience.

Foss, J.E. 1993. Subjectivity, objectivity, and Nagel on consciousness. *Dialogue* 32:725-36.

Nagel conflates metaphysical and epistemological versions of the subjective/objective distinction. Consciousness is metaphysically subjective, and science is epistemically objective, so there is incompatibility.

Francescotti, R.M. 1993. Subjective experience and points of view. *Journal of Philosophical Research* 18:25-36.

Being graspable from only one point of view does not define the class of

facts about conscious experience. Various ways of cashing this out fail.

Haksar, V. 1981. Nagel on subjective and objective. *Inquiry* 24:105-21.

The objective and subjective don't conflict, but complement each other.

Hanna, P. 1990. Must thinking bats be conscious? *Philosophical Investigations* 13:350-55.

Hiley, D.R. 1978. Materialism and the inner life. *Southern Journal of Philosophy* 16:61-70.

Nagel conflates questions about sensory qualities with those about a unique point of view. The truth of physicalism is irrelevant to uniqueness.

Hill, C.S. 1977. Of bats, brains, and minds. *Philosophy and Phenomenological Research* 38:100-106.

Kekes, J. 1977. Physicalism and subjectivity. *Philosophy and Phenomenological Research* 37:533-6.

The subjective/objective distinction is ill-drawn. Objective descriptions aren't species-independent, but in terms of the space-time causal network. Science can explain the experience this way, but not provide the experience.

Lewis, D. 1983. Postscript to "Mad pain and Martian pain". In *Philosophical Papers*, Vol. 1. Cambridge University Press.

Knowing what it's like consists in an ability, not possession of information.

Lycan, W.G. 1987. "Subjectivity". In *Consciousness*. MIT Press.

Various anti-Nagel points.

Lycan, W.G. 1990. What is the "subjectivity" of the mental? *Philosophical Perspectives*.

The subjectivity of the mental is no more special than usual propositional subjectivity. It can be handled by a self-scanner model of introspection.

Malcolm, N. 1988. Subjectivity. *Philosophy* 63:147-60.

A critique of Nagel's idea of a "point of view" that is occupied by a "subject". There aren't any peculiar facts about given viewpoints.

Maloney, J.C. 1986. About being a bat. *Australasian Journal of Philosophy* 64:26-49.

Mandik, P. 2001. Mental representation and the subjectivity of consciousness. *Philosophical Psychology* 14:179-202.

Mellor, D.H. 1993. Nothing like experience. *Proceedings of the Aristotelian Society* 63:1-16.

There are no facts about what an experience is like. Knowing what it's like is an ability to imagine, recognize, and recall; this explains ineffability, etc. With remarks on the experience of imagining an experience.

McClamrock, R. 1992. Irreducibility and subjectivity. *Philosophical Studies* 67:177-92.

Phenomenological properties cannot be picked out in physical or computational terms; argues against Lycan's criticism of Nagel. But all this is compatible with materialism. With comments on the phenomenological tradition.

McCulloch, G. 1988. What it is like. *Philosophical Quarterly* 38:1-19.

Criticizes absent/inverted qualia arguments for a special "what it is like", but argues that the possibility of "what it is like" differences relative to semantic states shows that something's not conveyed by functional accounts.

McMullen, C. 1985. 'Knowing what it's like' and the essential indexical. *Philosophical Studies* 48:211-33.

The Nagel/Jackson argument is analogous to the Perry indexical argument, and can be treated the same way.

Mounce, H.O. 1992. On Nagel and consciousness. *Philosophical Investigations* 15:178-84.

Muscari, P. 1985. The subjective character of experience. *Journal of Mind and Behavior* 6:577-97.

Muscari, P. 1987. The status of humans in Nagel's phenomenology. *Philosophical Forum* 19:23-33.

Nagel's dilemma: separating feeling from process. Moral consequences?

Nelkin, N. 1987. What is it like to be a person? *Mind and Language* 2:220-41.

Nagel-consciousness exists, but isn't so important. It's essential for sensations, but not for thoughts. Beings without it could still be persons.

Nemirow, L. 1980. Review of Nagel's *Mortal Questions*. *Philosophical Review* 89:473-7.

Understanding does not consist only in facts; we can understand via sympathy.

Nemirow, L. 1990. Physicalism and the cognitive role of acquaintance. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Knowing what it's like is really knowing how to imagine. We should reduce Nagel's question to a question about possession of a certain ability.

Pugmire, D. 1989. Bat or batman. *Philosophy* 64:207-17.

Subjectivity is not something we have knowledge of, as we lack comparisons.

Rorty, R. 1993. Holism, intrinsicity, and the ambition of transcendence. In (B. Dahlbom, ed) *Dennett and His Critics*. Blackwell.

On the Nagel/Dennett debate: Nagel holds out for unexplained intrinsic properties once the relational is all accounted for; Dennett can renounce the transcendental ambition. Remarks on realism, holism, and metaphilosophy.

Rudd, A.J. 1999. What it's like and what's really wrong with physicalism: A Wittgensteinian perspective. *Journal of Consciousness Studies* 5:454-63.

Russow, L. 1982. It's not like that to be a bat. *Behaviorism* 10:55-63.

Divides Nagel's problem: qualitative differences, special access, mineness.

Simoni-Wastila, H. 2000. Particularity and consciousness: Wittgenstein and Nagel on privacy, beetles and bats. *Philosophy Today* 44:415-425.

Taliaferro, C. 1988. Nagel's vista or taking subjectivity seriously. *Southern Journal of Philosophy* 26:393-401.

Nagel's 'View from Nowhere' doesn't take subjectivity seriously enough.

Teller, P. 1992. Subjectivity and knowing what it's like. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Rebutting various intuitions for the non-physical nature of experience. The Nagel/Jackson argument commits an intensional fallacy; experiences are physical states known from a different perspective.

Tilghman, B.R. 1991. What is it like to be an aardvark? *Philosophy* 66:325-38.

A Wittgensteinian critique of Nagel. Nagel's question is confused: "what it's like" is a matter of behavior, sociality, etc, not inner experience.

van Gulick, R. 1985. Physicalism and the subjectivity of the mental. *Philosophical Topics* 13:51-70.

Reducing doesn't imply understanding. Two different kinds of reduction.

Wider, K. 1989. Overtones of solipsism in Nagel's 'What is it like to be a bat?' and 'The view from nowhere'. *Philosophy and Phenomenological Research* 49:481-99.

Nagel is an epistemological solipsist, whether he likes it or not.

Wright, E. 1996. What it isn't like. *American Philosophical Quarterly* 33:23-42.

1.2b The Explanatory Gap (Levine)

Beckermann, A. 2000. The perennial problem of the reductive explainability of phenomenal consciousness: C.D. Broad on the explanatory gap. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Bieri, P. 1995. Why is consciousness puzzling? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Reflections on the explanatory gap between physical processes and conscious experience. With remarks on different sorts of consciousness, and on why we need an intelligible necessary connection.

Block, N. & Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review*.

Chalmers, D.J. & Jackson, F. 2001. Conceptual analysis and reductive explanation. *Philosophical Review*.

Ellis, R.D. & Newton, N. 1998. Three paradoxes of phenomenal consciousness: Bridging the explanatory gap. *Journal of Consciousness Studies* 5:419-42.

Hardin, C.L. 1987. Qualia and materialism: Closing the explanatory gap. *Philosophy and Phenomenological Research* 48:281-98.

On the physiological bases of phenomenal states, particularly color.

Inverted spectrum isn't really coherent, as coolness/warmth would have to be inverted too. So the contingency of qualia is diminished.

Hardin, C.L. 1992. Physiology, phenomenology, and Spinoza's true colors. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Argues in detail that psychophysics can provide a structural map to close the explanatory gap. If there is an explanatory residue, perhaps panpsychism can help.

Harnad, S. 1994. Why and how we are not zombies. *Journal of Consciousness Studies* 1:164-67.

Kim, J. 1998. Reduction, reductive explanation, and "the explanatory gap". Manuscript.

Argues for a distinction between reduction and reductive explanation, and argues that reductive explanations generally involves conceptual connections via "functionalization". With comments on Block and Stalnaker.

Levin, J. 1991. Analytic functionalism and the reduction of phenomenal states. *Philosophical Studies* 61:211-38.

Contra Kripkean arguments, a good enough functional theory may help close the conceivability/explanatory gap between the physical and qualia. Contra Nagel/Jackson, such a theory could provide us with recognitional abilities.

Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.

How do we explain the apparent contingency of the qualia-matter reduction? Even if it's not metaphysically contingent, it's conceptually contingent, so there's a gap in any physical explanation of qualia. Excellent.

Levine, J. 1993. On leaving out what it's like. In (M. Davies and G. Humphreys, eds) *_Consciousness: Psychological and Philosophical Essays_*. Blackwell.

Physical accounts leave out qualia epistemologically but not metaphysically. So physicalism holds, but there is an explanatory gap. Discusses Kripke's and Jackson's arguments in detail; also explanation and content.

Papineau, D. 1998. Mind the gap. *Philosophical Perspectives* 12:373-89.

Price, M.C. 1996. Should we expect to feel as if we understand consciousness? *Journal of Consciousness Studies* 3:303-12.

Argues that the explanatory gap between brain and consciousness is just the same as that found with causal relations everywhere; it's just that we usually overlook it in the latter case.

Sturgeon, S. 1994. The epistemic basis of subjectivity. *Journal of Philosophy* 91:221-35.

Qualia can't be explained in more basic terms, as that sort of explanation works by accounting for a property's canonical evidence, but the canonical evidence for qualia are qualia themselves. But they still may be physical.

Tye, M. 1999. Phenomenal consciousness: The explanatory gap as a cognitive illusion. *Mind* 108:705-25.

1.2c 'Hard' and 'Easy' Problems (Chalmers) [see also 3.1]

Chalmers, D.J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:200-19. Also in (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press. Reprinted in Shear 1997.

Divides the problems of consciousness into easy and hard problems; the hard problem eludes reductive explanation as it isn't about explaining functions. Argues instead for a nonreductive theory with psychophysical laws.

Chalmers, D.J. 1995. The puzzle of conscious experience. *Scientific American* 273(6):80-86.

Like the JCS article, but shorter, more accessible, and with pretty pictures.

Chalmers, D.J. 1996. Can consciousness be reductively explained? In *_The Conscious Mind_*. Oxford University Press.

There is no a priori entailment from the physical to phenomenal facts (arguments from conceivability, epistemology, analysis), so reductive explanation fails. With a critique of existing empirical proposals.

Chalmers, D.J. 1997. Moving forward on the problem of consciousness. *Journal of Consciousness Studies* 4:3-46. Reprinted in Shear 1997.

A reply to the 25 "hard problem" articles in JCS.

Chalmers, D.J. 1998. The problems of consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Churchland, P.S. 1996. The hornswoggle problem. *Journal of Consciousness Studies* 3:402-8. Reprinted in Shear 1997,

Argues that the "hard problem" in effect invokes an argument from ignorance, and that there's no deep difference between consciousness and other domains.

Clark, T. 1995. Function and phenomenology: Closing the explanatory gap. *Journal of Consciousness Studies* 2:241-54. Reprinted in Shear 1997.

Argues contra Chalmers that experience is identical to certain functions, rather than emerging from them.

- Crick, F. and Koch, C. 1995. Why neuroscience may be able to explain consciousness. *Scientific American* 273(6):84-85. Reprinted in Shear 1997.
Divides the hard problem into three parts, and argues that neuroscience can make progress on at least one part (incommunicability); and maybe "meaning" holds the key to the rest.
- Dennett, D.C. 1996. Facing backwards on the problem of consciousness. *Journal of Consciousness Studies* 3:4-6. Reprinted in Shear 1997.
Argues contra Chalmers 1995 that functions are all we need to explain.
- Eilan, N. 2000. Primitive consciousness and the 'hard problem'. *Journal of Consciousness Studies* 7:28-39.
- Gray, J. 1998. Creeping up on the hard question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Hodgson, D. 1996. The easy problems ain't so easy. *Journal of Consciousness Studies* 3:69-75. Reprinted in Shear 1997.
Argues that consciousness plays a vital role in performing mental functions, so the easy problems won't be solved until the hard problem is solved.
- Horst, S. 1999. Evolutionary explanation and the hard problem of consciousness. *Journal of Consciousness Studies* 6:39-48.
- Ismael, J. 1999. Science and the phenomenal. *Philosophy of Science* 66:351-69.
- Lewis, H. 1998. Consciousness: Inexplicable - and useless too? *Journal of Consciousness Studies* 5:59-66.
- Libet, B. 1996. Solutions to the hard problem of consciousness. *Journal of Consciousness Studies* 3:33-35. Reprinted in Shear 1997.
Endorses the idea of consciousness as fundamental, but criticizes Chalmers' psychophysical laws. Advocates a theory with a "conscious mental field".
- Lowe, E.J. 1995. There are no easy problems of consciousness. *Journal of Consciousness Studies* 2:266-71. Reprinted in Shear 1997.
Argues that the "easy problems" -- reportability, attention, etc -- all involve concepts and therefore experience itself, for Kantian reasons, and therefore are not mechanistically explainable.
- Mills, E.O. 1996. Giving up on the hard problem of consciousness. *Journal of Consciousness Studies* 3:26-32. Reprinted in Shear 1997.
Argues that the truly hard problem is that of giving a constitutive account of consciousness, and Chalmers doesn't solve that (laws aren't good enough); in fact it's unsolvable.
- Mills, F.B. 1998. The easy and hard problems of consciousness: A Cartesian perspective. *Journal of Mind and Behavior* 19:119-40.
- O'Hara, K. & Scutt, T. 1996. There is no hard problem of consciousness. *Journal of Consciousness Studies* 3:290-302. Reprinted in Shear 1997.
Argue that we should work on the easy problems for now, as nobody has any good ideas about the hard problem; maybe it will gradually fade away.
- Robinson, W.S. 1996. The hardness of the hard problem. *Journal of Consciousness Studies* 3:14-25. Reprinted in Shear 1997.
- Shear, J. 1996. The hard problem: Closing the empirical gap. *Journal of Consciousness Studies* 3:54-68. Reprinted in Shear 1997.
On the epistemology of the hard problem. Argues that a scientific study of phenomenology is possible, drawing on work in developmental psychology and Eastern thought. "Pure consciousness" may be relevant to a resolution.

Shear, J. (ed) 1997. Explaining Consciousness: The Hard Problem. MIT Press.

A collection of essays consisting of Chalmers' keynote paper, 26 replies from many perspectives, and Chalmers' response to the replies.

Varela, F. 1995. Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies* 3:330-49. Reprinted in Shear 1997. Advocates a careful phenomenological study of consciousness in its own right, systematically linked with a neurophysiological investigation.

Velmans, M. 1995. The relation of consciousness to the material world. *Journal of Consciousness Studies* 2:255-65. Reprinted in Shear 1997. Agrees with Chalmers on nonreductionism, but disagrees on "awareness", organizational invariance, and thermostats. Advocates a kind of dual-aspect theory, where the physical world is present within consciousness,

1.2d Cognitive Closure (McGinn)

Davies, W.M. 1999. Sir William Mitchell and the "new mysterianism". *Australasian Journal of Philosophy* 77:253-73.

Garvey, J. 1997. What does McGinn think we cannot know? *Analysis* 57:196-201.

Hanson, P.P. 1993. McGinn's cognitive closure. *Dialogue* 32:579-85.

Kirk, R. 1991. Why shouldn't we be able to solve the mind-body problem? *Analysis* 51:17-23.

McGinn asks too much of a solution to the M-B problem. We might understand consciousness without understanding specific experiences; we could get at it by studying brain and consciousness not separately but simultaneously.

Krellenstein, M.F. 1995. Unsolvable problems, visual imagery, and explanatory satisfaction. *Journal of Mind and Behavior* 16:235-54.

Kukla, A. 1995. Mystery, mind, and materialism. *Philosophical Psychology* 8:255-64.

McDonough, R. 1992. The last stand of mechanism. *Journal of Speculative Philosophy* 6:206-25.

McGinn, C. 1989. Can we solve the mind-body problem? *Mind* 98:349-66. Reprinted in The Problem of Consciousness (Blackwell, 1991).

Argues that the mind-body problem might be solvable in principle, but beyond human capacities. Neither perception of the brain nor introspection of consciousness can uncover the property by which consciousness arises.

McGinn, C. 1991. The Problem of Consciousness: Essays Toward a Resolution. Blackwell.

A collection of articles on the problem of consciousness, advocating a view on which the phenomenon is natural but permanently mysterious to us.

McGinn, C. 1991. Consciousness and the natural order. In The Problem of Consciousness. Blackwell.

Argues that a naturalistic account of the intentionality of conscious states requires an account of their embodiment; and that embodiment may depend on the hidden structure of conscious states, not accessible to introspection.

McGinn, C. 1991. The hidden structure of consciousness. In The Problem of Consciousness. Blackwell.

Suggests that consciousness may have a hidden structure, analogous to the deep structure of language, that relates its surface properties to physical

properties. We may not be able to understand this hidden structure, however.

McGinn, C. 1993. *Problems in Philosophy*. Blackwell.

McGinn, C. 1995. Consciousness and space. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

McGinn, C. 1999. *The Mysterious Flame: Conscious Minds in a Material World*. Basic Books.

Sacks, M. 1994. Cognitive closure and the limits of understanding. *Ratio* 7:26-42

Whitely, C.H. 1990. McGinn on the mind-body problem. *Mind* 99:289.

1.2e Miscellaneous

Churchland, P.M. 1996. The rediscovery of light. *Journal of Philosophy* 93:211-28.

Parodies arguments by Searle, Jackson, and Chalmers for the irreducibility of consciousness with analogous arguments for the irreducibility of "luminescence". The consciousness arguments are no better.

Churchland, P.S. 1998. What should we expect from a theory of consciousness? In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Hardcastle, V.G. 1993. The naturalists versus the skeptics: The debate over a scientific understanding of consciousness. *Journal of Mind and Behavior* 14:27-50.

Argues that consciousness can be handled within a scientific framework. We can translate first-person accounts into third-person accounts. Replies to skeptical objections using analogies from elsewhere in science.

Hardcastle, V.G. 1996. The why of consciousness: A non-issue for materialists. *Journal of Consciousness Studies* 3:7-13.

A "committed materialist" will not see any explanatory gap, or any "brute fact". The entrenched differences lie in one's choice of initial framework.

Hesslow, G. 1996. Will neuroscience explain consciousness? *Journal of Theoretical Biology* 171:29-39.

Kirk, R. 1995. How is consciousness possible? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Kurthen, M. 1995. On the prospects of a naturalistic theory of phenomenal consciousness. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Lockwood, M. 1998. The enigma of sentience. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Nida-Rumelin, M. 1997. Is the naturalization of qualitative experience possible or sensible? In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Taylor, J.G, 1998. Cortical activity and the explanatory gap. *Consciousness and Cognition* 7:109-48.

van Gulick, R. 1993. Understanding the phenomenal mind: Are we all just armadillos? In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Qualia pose no insurmountable problems for materialism: knowledge argument can be answered, explanatory gap can be closed, and absent qualia arguments beg the question. With speculations on their functional role.

van Gulick, R. 1995. What would count as explaining consciousness? In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Distinguishes six explananda, four explanatory restrictions, and four sorts of relations between them, making 96 possible problems. With a discussion of whether and how the central problems might be answered.

1.3 Materialism and Dualism

1.3a The Knowledge Argument (Jackson) [see also 1.2a]

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Knowing a completed neuroscience does not imply knowing about qualia. Mary, the colorblind neuroscientist, gains color vision and learns about red. So physicalism is false, as there are facts over and above the physical facts.

Jackson, F. 1986. What Mary didn't know. *Journal of Philosophy* 83:291-5.

Reply to Churchland 1985: Mary **learns**, Churchland misstates the argument.

Alter, T. 1995. Mary's new perspective. *Australasian Journal of Philosophy* 73:585-84.

Contra Pereboom 1994: The way a color sensation appears is a fact about it.

Alter, T. 1998. A limited defense of the knowledge argument. *Philosophical Studies* 90:35-56.

Bachrach, J.E. 1990. Qualia and theory reduction: A criticism of Paul Churchland. *Iyyun* 281-94.

Argues that Churchland's neuroscientific descriptions must leave at least some qualia behind: they might account for what we know (e.g. brain states) in qualia-knowledge, but can't handle distinctions in how we know.

Bigelow, J. & Pargetter, R. 1990. Acquaintance with qualia. *Theoria*.

Mary gains knowledge of old facts, in a new way: she gains a new mode of acquaintance with those facts. Analogies with indexical knowledge: her new knowledge eliminates no possible worlds.

Churchland, P.M. 1985. Reduction, qualia and the direct introspection of brain states. *Journal of Philosophy* 82:8-28. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).

Qualia can undergo a normal reduction to the neurophysiological. Jackson commits an intensional fallacy; in any case, perhaps Mary can understand red. When we apprehend qualia, we are directly introspecting our brain state.

Churchland, P.M. 1989. Knowing qualia: A reply to Jackson. In *_A Neurocomputational Perspective_*. MIT Press.

Rejoinder to Jackson 1986. The key lies in knowing-how vs. knowing-that.

Conee, E. 1985. Physicalism and phenomenal properties. *Philosophical Quarterly* 35:296-302.

Contra Lewis, Nemirow, and Horgan on the knowledge argument. But qualia may still be physical (though outside vocab of science) due to their causal role.

Conee, E. 1994. Phenomenal knowledge. *Australasian Journal of Philosophy*.

Mary knew all the facts about qualia beforehand, she just wasn't acquainted with them. Knowledge by acquaintance isn't factual knowledge.

Cummins, R. 1984. The mind of the matter: Comments on Paul Churchland.
Philosophy of Science Association 1984, 2:791-8.

Speculation on how consciousness might be left out by a physical account.

Dennett, D.C. 1991. "Epiphenomenal" qualia? In *_Consciousness Explained_*, pp.
398-406. Little-Brown.

Argues that most people don't really imagine Mary's situation. In fact,
Mary would be able to identify blue objects from the way they make her react.

Furash, G. 1989. Frank Jackson's knowledge argument against materialism.
Dialogue 32:1-6.

Defends Jackson's argument against criticisms by Nemirow, Smith & Jones,
Warner, Horgan, & Conee. The argument forces physicalism into a quandary:
either deny qualia, or make the confused claim that qualia are physical.

Gertler, B. 1999. A defense of the knowledge argument. *Philosophical Studies*
93:317-336.

Graham, G. & Horgan, T. 2000. Mary Mary, quite contrary. *Philosophical*
Studies 99:59-87.

Harman, G. 1993. Can science understand the mind? In (G. Harman, ed)
Conceptions of the Human Mind: Essays on Honor of George A. Miller. Lawrence
Erlbaum.

On Dilthey's "Verstehen", or "understanding from within". Mostly about
meaning, but with application to the knowledge Mary gains.

Hershfield, J. 1998. Lycan on the subjectivity of the mental. *Philosophical*
Psychology 11:229-38.

Horgan, T. 1984. Jackson on physical information and qualia. *Philosophical*
Quarterly 34:147-83.

Mary didn't know all the physical facts: she knew all the explicitly physical
information, but not all the ontologically physical information.

Jacquette, D. 1995. The blue banana trick: Dennett on Jackson's color
scientist. *Theoria* 61:217-30.

Kelly, J.S. 1989. On neutralizing introspection: The data of sensuous
awareness. *Southern Journal of Philosophy* 27:29-53.

Lahav, R. 1994. A new challenge for the physicalist: Phenomenal
indistinguishability. *Philosophia* 24:77-103.

A new version of the knowledge argument: given all the physical facts, one
can't know when two experiences are indistinguishable. This avoids
various objections to the standard version.

Levin, J. 1986. Could love be like a heatwave?: Physicalism and the subjective
character of experience. *Philosophical Studies* 49:245-61. Reprinted in (W.
Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Contra Nagel/Jackson: Understand qualia through relational properties, and
separate the mental concept from the recognitional capacity.

Lewis, D. 1990. What experience teaches. In (W. Lycan, ed) *_Mind and*
Cognition_. Blackwell.

Against the hypothesis that phenomenology carries information. If it does,
then qualia are epiphenomenal. Better to analyze the "new information" as
acquiring an ability instead. In-depth and entertaining.

Loar, B. 1990. Phenomenal states. *Philosophical Perspectives* 4:81-108.

Phenomenal and functional concepts are distinct, but the relevant properties
may be identical. We directly refer to phenomenal properties by recognition.
Remarks on other minds, transparency, incorrigibility & more. A meaty paper.

Lycan, W.G. 1995. A limited defense of phenomenal information. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.
Gives nine arguments against the Lewis/Nemirow ability analysis, and proposes instead that the same fact is learned a new way, like water and H₂O facts.
This sort of phenomenal information is no danger to materialism.

Lycan, W.G. 1998. Phenomenal information again: It is both real and intrinsically perspectival. *Philosophical Psychology* 11:239-42.

McConnell, J. 1995. In defense of the knowledge argument. *Philosophical Topics* 22:157-187.
Defends against objections from Dennett, Churchland, etc. Horgan's objection (same fact different ways) has a certain force, but the argument can be reformulated to avoid them and imply property dualism. With remarks on Loar.

Nemirow, L. 1995. Understanding rules. *Journal of Philosophy* 92:28-43.

Newton, N. 1986. Churchland on direct introspection of brain states. *Analysis* 46:97-102.
Contra Churchland 1985: we couldn't introspect sensations as brain states, although we could interpret them as such.

Nida-Rumelin, M. 1995. What Mary couldn't know: Belief about phenomenal states. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.
Reconstructing the Mary case as a Marianna case, and introducing a distinction between phenomenal and nonphenomenal beliefs, to support the knowledge argument. A very nice and sophisticated paper.

Nida-Rumelin, M. 1998. On belief about experiences: An epistemological distinction applied to the knowledge argument against physicalism. *Philosophy and Phenomenological Research* 58:51-73.

Papineau, D. 1993. Physicalism, consciousness, and the antipathetic fallacy. *Australasian Journal of Philosophy* 71:169-83.
Mary goes from a third-person concept of experience to a first-person concept, but they co-refer; we can refer to an experience without having the experience. Physical and phenomenal properties are brutally identical.

Papineau, D. 1995. The antipathetic fallacy and the boundaries of consciousness. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Pereboom, D. 1994. Bats, brain scientists, and the limits of introspection. *Philosophy and Phenomenological Research* 54:315-29.
Mary learns an old fact under a new mode of presentation, and doesn't even learn a new fact about a mode of presentation. Her access to internal states is always mediated by representation, so we can always ascend to a new mode.

Raymont, P. 1995. Tye's criticism of the knowledge argument. *Dialogue* 34:713-26.

Raymont, P. 1999. The know-how response to Jackson's knowledge argument. *Journal of Philosophical Research* 24:113-26.

Robinson, D. 1993. Epiphenomenalism, laws, and properties. *Philosophical Studies* 69:1-34.
A thorough discussion of Lewis 1990. Phenomenal information implies epiphenomenalism, even at the intra-psychic level. Remarks on ineffability, and on whether properties should be individuated by nomic role or by essence.

Robinson, H. 1993. Dennett on the knowledge argument. *Analysis* 53:174-7.
Contra Dennett, Mary can't tell an object's color unless she already knows

about experience. The knowledge argument bears on thought, not just qualia.

Robinson, H. 1993. The anti-materialist strategy and the "knowledge argument". In (H. Robinson, ed) *_Objections to Physicalism_*. Oxford University Press.

Shoemaker, S. 1984. Churchland on reduction, qualia, and introspection. *Philosophy of Science Association* 1984, 2:799-809.

Introspection reveals functional properties, not physical, so qualia should be reduced to the functional, not to the physical.

Stemmer, N. 1989. Physicalism and the argument from knowledge. *Australasian Journal of Philosophy* 67:84-91.

Physicalism explains all the relevant evidence, hence all facts, and needn't admit mental entities; belief in mental entities is based on physical facts.

Thompson, E. 1992. Novel colors. *Philosophical Studies* 68:321-49.

Interesting remarks on what it would be for someone to see colors that we cannot, combining philosophical considerations with empirical findings about color space. Argues that science could tell us what such colors are like.

Tye, M. 2000. Knowing what it is like: The ability hypothesis and the knowledge argument. In *_Consciousness, Color, and Content_*. MIT Press.

Warner, R. 1986. A challenge to physicalism. *Australasian Journal of Philosophy* 64:249-65.

A Jackson-like argument that physical knowledge can't give you the knowledge of what pain feels like. With detailed consideration of objections and replies. Argues from limited incorrigibility to factualism about pains.

Watkins, M. 1989. The knowledge argument against the knowledge argument. *Analysis*, 49:158-60.

Epiphenomenalism => qualia don't cause beliefs => we don't know about qualia.

Zemach, E. 1990. Churchland, introspection, and dualism. *Philosophia* 20:3-13.

1.3b Zombies & Modal Arguments [see also 1.2c, 1.3a, 1.3c, 1.3d, 1.7e]

Balog, K. 1999. Conceivability, possibility, and the mind-body problem. *Philosophical Review* 108:497-528.

Bringsjord, S. 1999. The zombie attack on the computational conception of mind. *Philosophy and Phenomenological Research* 59:41-69.

Brueckner, A. 2001. Chalmers' conceivability argument for dualism. *Analysis*.

Byrne, A. 1999. Cosmic hermeneutics. *Philosophical Perspectives*.

Chalmers, D.J. 1996. Naturalistic dualism. In *_The Conscious Mind_*. Oxford University Press.

Argues from the lack of logical supervenience to the falsity of physicalism. A two-dimensional analysis shows that objections from a posteriori necessity fail. Argues for a naturalistic variety of property dualism.

Chalmers, D.J. 2002. Does conceivability entail possibility? In (T. Gendler & J. Hawthorne, eds) *_Imagination, Conceivability, and Possibility_*. Oxford University Press.

Cottrell, A. 1999. Sniffing the camembert: On the conceivability of zombies. *Journal of Consciousness Studies* 6:4-12.

Dennett, D.C. 1995. The unimagined preposterousness of zombies. *Journal of Consciousness Studies* 2:322-26.

Guzeldere, G. 1995. Varieties of zombiehood. *Journal of Consciousness Studies* 2:326-33.

Hill, C.S. 1997. Imaginability, conceivability, possibility, and the mind-body problem. *Philosophical Studies* 87:61-85.

Argues that the conceivability of zombies and the like can be explained away, in terms of the cognitive separability of perceptual imagination and sympathetic imagination of the same states.

Hill, C.S. 1998. Chalmers on the apriority of modal knowledge. *Analysis* 58:20-26.

Hill, C.S. & McLaughlin, B.P. 1998. There are fewer things in reality than are dreamt of in Chalmers' philosophy. *Philosophy and Phenomenological Research*.

Kirk, R. 1974. Sentience and behaviour. *Mind* 81:43-60.

Describing a situation where we would be justified in believing in zombies. Argues that zombies are logically possible, which seems incompatible with most or all varieties of materialism.

Kirk, R. 1974. Zombies vs materialists. *Aristotelian Society Supplement* 48:135-52.

Materialism requires that physical states logically entail all non-relational states; but zombies are logically possible, so materialism fails. With a description of a zombie, and replies to a verificationist. All very true.

Kirk, R. 1977. Reply to Don Locke on zombies and materialism. *Mind* 86:262-4.

Reply to Locke 1976: materialism needs zombies to be logically impossible.

Kirk, R. 1999. Why there couldn't be zombies. *Proceedings of the Aristotelian Society, Supplementary Volume* 73:1-16.

Kraemer, E.R. 1980. Imitation-man and the 'new' epiphenomenalism. *Canadian Journal of Philosophy* 10:479-487.

If Campbell's imitation man is possible, then the causal relation between the physical and phenomenal is unreliable.

Latham, N. 1998. Chalmers on the addition of consciousness to the physical world. *Philosophical Studies*.

Levine, J. 1998. Conceivability and the metaphysics of mind. *Nous* 32:449-480.

Locke, D. 1976. Zombies, schizophrenics, and purely physical objects. *Mind* 83:97-99.

Contra Kirk: the logical possibility of zombies is compatible with empirical materialism. With some comments on Kirk's thought-experiment.

Marton, P. 1998. Zombies vs. materialists: The battle over conceivability. *Southwest Philosophy Review* 14:131-38.

Melnyk, A. 1998. Physicalism unfalsified: Chalmers' inconclusive argument for dualism. In (B. Loewer & C. Gillett, eds) *Physicalism and its Discontents*. Oxford University Press.

Moody, T. 1994. Conversations with zombies. *Journal of Consciousness Studies* 1:196-200.

Argues that behavioral differences in zombies would show up, in their discourse about consciousness.

Nagel, T. 1998. Conceiving the impossible and the mind-body problem. *Philosophy* 73:337-52.

Perkins, M. 1970. Matter, sensation, and understanding. *American Philosophical Quarterly* 8:1-12.

On the possibility of an Insentient Perceiver, who perceives the world without sensation. Sensation is inessential to perception and understanding, except understanding in the "whatlike" manner.

Perkins, M. 1971. Sentience. *Journal of Philosophy* 68:329-37.

Argues for the conceivability of insentient perception of colors (in "Insent", a kind of blindsighter or zombie), in order to argue for a realistic account of colors.

Perry, J. 2001. The zombie argument. In *_Knowledge, Possibility, and Consciousness_*. MIT Press.

Perry, J. 2001. The modal argument. In *_Knowledge, Possibility, and Consciousness_*. MIT Press.

Prudovsky, G. 1995. Arguments from conceivability. *Ratio* 8:63-69.

Robb, D. 1999. Conceivability and consciousness. *Philosophical Topics*.

Robinson, H. 1976. The mind-body problem in contemporary philosophy. *Zygon* 11:346-360.

A discussion of materialism and its difficulties. The conceivability of zombies poses special problems. Criticism of Smart's & Armstrong's analyses.

Squires, R. 1974. Zombies vs materialists II. *Aristotelian Society Supplement* 48:153-63.

Stalnaker, R. 2002. What is it like to be a zombie? In (T. Gendler & J. Hawthorne, eds) *_Imagination, Coceivability, and Possibility_*. Oxford University Press.

Stoljar, D. 2000. Physicalism and the necessary a posteriori. *Journal of Philosophy* 87:33-55.

Thomas, N. 1998. Zombie killer. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

van Gulick, R. 1999. Conceiving beyond our means: The limits of thought experiments. In (S. Hameroff, A. Kaszniak, & D. Chalmers, eds) *_Toward a Science of Consciousness III_*. MIT Press.

Yablo, S. 1998. Concepts and consciousness. *Philosophy and Phenomenological Research*.

Yablo, S. 1998. Textbook Kripkeanism and the open texture of language. *Philosophical Quarterly*.

1.3c Essentialist Arguments (Kripke) [see also 1.3b, 1.3d]

Kripke, S.A. 1971. Identity and necessity. In (M. Munitz, ed) *_Identity and Individuation_*.

An identity between mental and physical states can't be contingent, as it relates rigid designators. But nevertheless the co-occurrence of certain mental and physical states is contingent, so the identity theory is false.

Kripke, S.A. 1972. *_Naming and Necessity_*. Harvard University Press.

Both "pain" and "C-fibres firing" are rigid designators, so if they are identical, this must be necessary. But their co-occurrence is contingent, and this can't be explained away epistemically, so the identity theory fails.

- Barnette, R. 1977. Kripke's pains. *Southern Journal of Philosophy* 15.
Argues that pain and the associated epistemic situation are inequivalent.
Beliefs about pain are simply produced by mechanisms, and could come about without any sensation.
- Bayne, S.R. 19xx. Kripke's Cartesian argument. *Philosophia*.
Trying to turn Kripke's argument against him: it's possible that pains and C-fibre stimulations are identical, so it's necessary that they're identical.
- Bealer, G. 1994. Mental properties. *Journal of Philosophy* 91:185-208.
On four arguments against the identity theory: multiple-realizability, modal, knowledge, and certainty arguments. All face difficulties due to scientific essentialism, but the latter two can be reformulated to avoid them.
- Blumenfeld, J. 1975. Kripke's refutation of materialism. *Australasian Journal of Philosophy* 53:151-6.
Kripke's argument doesn't refute token identity. Pains can have other essential properties besides painfulness, so psychophysical token identities can be necessary.
- Boyd, R. 1980. Materialism without reductionism: What physicalism does not entail. In (N. Block, ed) *_Readings in the Philosophy of Psychology_*, vol 1. Harvard University Press.
Materialism doesn't need rigid identities, due to the compositional plasticity of mental states. So the possibility of disembodiment is compatible with materialism. The possibility of zombies is illusory.
- Carney, J. & von Bretzel, P. 1973. Modern materialism and essentialism. *Australasian Journal of Philosophy* 51:78-81.
A materialist must deny essentialism to meet Kripke's argument.
- Carney, J. 1975. Kripke and materialism. *Philosophical Studies* 27:279-282.
Comments on Feldman 1974: Feldman's view requires rejection of Kripke's views on necessity, or a problematic mixed view on rigid designators.
- Della Rocca, M. 1993. Kripke's essentialist arguments against the identity theory. *Philosophical Studies* 69:101-112.
Kripke's premise that pains are essentially mental either begs the question (by assuming pains don't have physical properties) or weakens the premise that physical events aren't essentially mental.
- Double, R. 1976. The inconclusiveness of Kripke's argument against the identity theory. *Auslegung* 3:156-65.
- Feldman, F. 1973. Kripke's argument against materialism. *Philosophical Studies* 24:416-19.
Painfulness need not be an essential feature of pains.
- Feldman, F. 1974. Kripke on the identity theory. *Journal of Philosophy* 71:665-76.
Kripke's arguments against person-body and mind-brain identity rely on the essentialness of aliveness to persons and painfulness to pains. There's no reason to grant this. If we do, rigidity is irrelevant to the argument.
- Feldman, F. 1980. Identity, necessity, and events. In (N. Block, ed) *_Readings in the Philosophy of Psychology_*, Vol. 1. Harvard University Press.
Defending a contingent event identity thesis against Kripke. Mental properties (which are distinct from physical properties) may not be essential properties of an event.
- Gjelsvik, O. 1988. A Kripkean objection to Kripke's arguments against the identity-theories. *Inquiry* 30:435-50.

Uses Kripke's 1979 direct-reference theory against him. When rigid designators don't have associated reference-fixing descriptions, we can't expect the "explaining away" strategy to work.

Hill, C.S. 1981. Why Cartesian intuitions are compatible with the identity thesis. *Philosophy and Phenomenological Research* 42:254-65.

The apparent contingency of identity is due to the fact that one can be aware of pain without being aware of C-fibers and vice versa, as well as to the fact that "C-fibers" may be picked out by a contingent description.

Holman, E. 1988. Qualia, Kripkean arguments, and subjectivity. *Philosophy Research Archives* 13:411-29.

Defending Kripkean arguments against various objections. Analysis in terms of manifest properties and their role in fixing reference to the subjective and objective.

Jackson, F. 1980. A note on physicalism and heat. *Australasian Journal of Philosophy* 58:26-34.

A Kripkean argument against non-analytic physicalism. Even if pain rigidly designates a brain state, the physicalist still has problems explaining the property of "pain-presents".

Jacquette, D. 1987. Kripke and the mind-body problem. *Dialectica* 41:293-300.

Kripke's argument doesn't refute contingent identity between minds and nonrigidly designated bodies, which is all materialism needs.

Leplin, J. 1979. Theoretical identification and the mind-body problem. *Philosophia* 8:673-88.

Some theoretical identifications are analogous to mental-physical identifications -- entities are introduced by properties considered essential within a theory, but this doesn't preclude identification.

Levin, M. 1975. Kripke's argument against the identity thesis. *Journal of Philosophy* 72:149-67.

The reference of "pain" is fixed not by essential features but by contingent topic-neutral descriptions; this is the real moral of Wittgenstein's private language argument. So Kripke's apparent contingency can be explained away?

Levin, M. 1995. Tortuous dualism. *Journal of Philosophy* 92:313-22.

Reply to Bealer 1994. Tries to clarify the dialectic, and argues that the materialist can explain "possibility" of straw thought as thought conjoined with mere appearance of straw.

Lycan, W.G. 1974. Kripke and the materialists. *Journal of Philosophy* 71:677-89.

Kripke equivocates on "pain-sensation": pains aren't the same as impressions of pain. Argues that imaginability arguments aren't decisive, and that functionalism may be less vulnerable than the identity theory.

Lycan, W.G. 1987. Functionalism and essence. In *_Consciousness_*. MIT Press.

Painfulness needn't be essential to pains: pains are events, not objects, and events don't have essences; and the reference of "pain" is fixed by topic-neutral descriptions. With remarks on pains vs. pain-sensations.

Maxwell, G. 1979. Rigid designators and mind-brain identity. *Minnesota Studies in the Philosophy of Science* 9.

"Brain state" reference is fixed by topic-neutral description; picks out pain but mightn't have, explaining the illusion of contingency. "Nonphysicalist materialism" results, with mind the essence of matter. An important paper.

McGinn, C. 1977. Anomalous monism and Kripke's Cartesian intuitions. *Analysis* 2:78-80. Reprinted in (N. Block, ed) *_Readings in the Philosophy of*

Psychology_ (MIT Press, 1980).

Token identity theories aren't vulnerable to Kripke's argument: it may be essential to this pain that it is a C-fibre firing, although not to pain as a type.

McMullen, C. 1984. An argument against the identity theory. Pacific Philosophical Quarterly 65:277-87.

We can explain away the apparent contingency of identity in terms of possible differences in *evidence* for the physical state. With a discussion on identities between something perceived and something described.

Mucciolo, L. 1975. On Kripke's argument against the identity thesis. Philosophia 5:499-506.

"Pain" need not be a rigid designator, but instead may pick out a state by its causal role. If it is a rigid designator, then the apparent contingency of identity comes from imagining something else filling the causal role.

Sher, G. 1977. Kripke, Cartesian intuitions, and materialism. Canadian Journal of Philosophy 7:227-38.

The reference of "C-fibre stimulation" might be fixed contingently, allowing the intuitive contingency of identity to be explained away.

Taylor, P. 1983. McGinn, token physicalism, and a rejoinder of Woodfield. Analysis 43:80-83.

Woodfield, A. 1978. Identity theories and the argument from epistemic counterparts. Analysis 38:140-3.

Contra McGinn 1977, the counterpart strategy fails as any pain that occurred here now would have been this pain. A counterpart strategy on brain states may work. With a reply by McGinn and a later rejoinder by Woodfield.

1.3d Arguments from Disembodiment [see also 1.3b, 1.3c]

Alston, W.P. & Smythe, T.W. 1994. Swinburne's argument for dualism. Faith and Philosophy 11:127-33.

Carrier, L. 1974. Definitions and disembodied minds. Personalist Forum 55:334-43.

Cole, D.J. & Foelber, F. 1984. Contingent materialism. Pacific Philosophical Quarterly 65:74-85.

Argues that materialism is only contingently true, as it's conceptually possible that we could become immaterial by gradual replacement.

Hart, W.D. 1988. _The Engines of the Soul_. Cambridge University Press.

Lewy, C. 1943. Is the notion of disembodied existence self-contradictory? Proceedings of the Aristotelian Society 43:59-78.

Long, D. 1977. Disembodied existence, physicalism, and the mind-body problem. Philosophical Studies.

Merricks, T. 1994. A new objection to a priori arguments for dualism. American Philosophical Quarterly 31:81-85.

Physicalism is compatible with the possibility of disembodiment: one can hold that mind and body are identical, that the body is physical, but that it is not essentially physical.

Odegard, D. 1970. Disembodied existence and central state materialism. Australasian Journal of Philosophy 48:256-60.

Pecnjak, D. 1995. Remarks on disembodied existence. Acta Analytica 10:209-13.

Shoemaker, S. 19xx On an argument for dualism. Reprinted in *_Identity, Cause, and Mind_*. Cambridge University Press, 1984.

Swinburne, R. 1997. The modal argument for substance dualism. In *_The Evolution of the Soul_* (revised edition). Oxford University Press.

Taliaferro, C. 1986. A modal argument for dualism. *Southern Journal of Philosophy* 24:95-108.

Taliaferro, C. 1997. Possibilities in the philosophy of mind. *Philosophy and Phenomenological Research* 57:127-37.

Tidman, P. 1994. Conceivability as a test for possibility. *American Philosophical Quarterly* 31:297-309.

Tye, M. 1983. On the possibility of disembodied existence. *Australasian Journal of Philosophy* 61:275-282.

There's no reason to believe that disembodied existence is possible: lack of logical contradiction doesn't imply possibility, conceivability is too weak a criterion, and it's not obvious that the situation is imaginable.

van Cleve, J. 1983. Conceivability and the Cartesian argument for dualism. *Pacific Philosophical Quarterly*.

Yablo, S. 1993. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53:1-42.

Zimmerman, D. 1991. Two Cartesian arguments for the simplicity of the soul. *American Philosophical Quarterly* 28:127-37.

1.3e Consciousness and Physicalism, Misc [see also 1.2, 1.3, 1.7b, 3.5b]

Fox, M. 1978. Beyond materialism. *Dialogue* 17:367-70.

Hill, C.S. 1991. *_Sensations: A Defense of Type Materialism_*. Cambridge University Press.

Defending type materialism, by way of criticism of dualism and functionalism. With treatments of introspection, sensory concepts, and other minds.

Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, & S. White, eds) *_Philosophy and the Cognitive Sciences_*. Holder-Pichler-Tempsky.

On why materialism requires conceptual analysis to locate mental properties in the natural world. Even a posteriori necessary connections have to be backed by a priori links. With remarks on supervenience. A nice paper.

Kirk, R. 1979. From physical explicability to full-blooded materialism. *Philosophical Quarterly* 29:229-37.

If every physical events has a physical explanation, and the mental is causally efficacious, then mental facts are strictly implied by physical facts. A nice argument.

Kirk, R. 1982. Physicalism, identity, and strict implication. *Ratio* 24:131-41.

Physicalism requires that all mental facts be strictly implied by the physical facts. Once this is recognized, questions about necessary or contingent identity are beside the point, and indeed identity is irrelevant.

Locke, D. 1971. Must a materialist pretend he's anaesthetized? *Philosophical Quarterly* 21:217-31.

Lund, D. 2000. Materialism and the subject of consciousness. *Idealistic Studies* 30:7-23.

Madell, G. 1988. *Mind and Materialism*. Edinburgh University Press.
On the problems posed for materialism by intentionality, autonomy, awareness, and indexicality. Tentatively advocates a Cartesian position.

Robinson, H. 1982. *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Robinson, H. (ed) 1993. *Objections to Physicalism*. Oxford University Press.
A collection of arguments against physicalism, mostly based on worries about consciousness and qualia.

Robinson, W.S. 1982. Sellarsian materialism. *Philosophy of Science* 49:212-27.

Seager, W.E. 1992. *Metaphysics of Consciousness*. Routledge and Kegan Paul.
Consciousness could be physical even if not explicable; but supervenience worries make it hard to see how it *could* be physical, though causal role suggests that it must be. We need a new conception. A stimulating book.

Sellars, W. 1981. Is consciousness physical? *Monist* 64:66-90.
On the place of "occurrent pink" and the "sensorium" in the physical world. It may turn out that the physics of the brain differs from other physics, in order to accommodate the causal role of sensations.

Smith, A.D. 1993. Non-reductive physicalism? In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.
A careful discussion of how to characterize physicalism, in terms of identity or supervenience, and argues that physicalism must reduce (bowdlerize) qualia to something they are not, as physicalism requires topic-neutral analyses.

1.3f Consciousness and Dualism [see also 1.2c, 1.3, 3.4d]

Collins, C. 1997. Searle on consciousness and dualism. *International Journal of Philosophical Studies* 5:15-33.

Double, R. 1983. Nagel's argument that mental properties are nonphysical. *Philosophy Research Archives* 9:217-22.

Eccles, J. 1987. Brain and mind: Two or one? In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Foster, J. 1989. A defense of dualism. In (J. Smythies & J. Beloff, eds) *The Case for Dualism*. University of Virginia Press.
Argues that all forms of materialism fail, and that dualism is the only option. Defends dualism against objections, and argues for interactionism over epiphenomenalism. A very clear and interesting paper.

Honderich, T. 1981. Psychophysical law-like connections and their problems. *Inquiry* 24:277-303.
Defending lawlike connections between physical states & conscious occurrents. Contra anomalous monism and identity theory for occurrents. But occurrents may not be causally efficacious. Comments by Wilson/Sprigge/Mackie/Stich.

Lahav, R. & Shanks, N. 1992. How to be a scientifically respectable 'property dualist'. *Journal of Mind and Behavior* 13:211-32.
Argues that the hypothesis of consciousness as an irreducible global property of the brain is compatible with what we know of both neuroscience and physics. With interesting remarks on quantum mechanics.

Latham, N. 2000. Chalmers on the addition of consciousness to the physical

world. Philosophical Studies 98:67-93.

McGinn, C. 1993. Consciousness and cosmology: Hyperdualism ventilated. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

A dialogue with a "hyperdualist". On the pros and cons of materialist vs. dualist ontology and cosmology. Dualism avoids the "magic" of emergence at the cost of an inflated and bizarre ontology.

O'Leary-Hawthorne, J. & McDonough, J.K. 1998. Numbers, minds, and bodies: A fresh look at mind-body dualism. *Philosophical Perspectives* 12:349-71.

Rosenberg, J.F. 1988. On not knowing what or who one is: Reflections on the intelligibility of dualism. *Topoi* 7:57-63.

Smook, R. 1988. Egoicity and twins. *Dialogue* 27:277-86.

Smythies, J.R. & Beloff, J. (eds) 1989. *The Case for Dualism*. University of Virginia Press.

Sprigge, T.L.S. 1994. Consciousness. *Synthese* 98:73-93.

On the non-physical nature of consciousness, and the threat of a merely contingent connection to behavior; suggests a denial of "Hume's principle". Perhaps consciousness is the noumenal essence of the physical. A nice paper.

Swinburne, R. 1986. *The Evolution of the Soul*. Oxford University Press.

Taliaferro, C. 1996. *Consciousness and the Mind of God*. Cambridge University Press.

von Wright, G.H. 1994. On mind and matter. *Journal of Theoretical Biology* 171:101-10.

1.3g Mind-Body Problem, General

Beck, L.W. 1940. The psychophysical as a pseudo-problem. *Journal of Philosophy* 37:561-71.

Butler, C.W. 1972. The mind-body problem: A nonmaterialistic identity thesis. *Idealistic Studies* 2:229-48.

Campbell, K.K. 1970. *Body and Mind*. Doubleday.

Carrier, M. & Mittelstrass, J. 1991. *Mind, Brain, Behavior: The Mind-Body Problem and the Philosophy of Psychology*. de Gruyter.

Cheng, C. (ed) 1975. *Philosophical Aspects of the Mind-Body Problem*. Hawaii University Press.

Cooper, W.E. 1977. Beyond materialism and back again. *Dialogue* 16:191-206.

Diaz, J. 2000. Mind-body unity, dual aspect, and the emergence of consciousness. *Philosophical Psychology* 13:393-403.

Feigl, H. 1934. Logical analysis of the psychophysical problem. *Philosophy of Science* 1:420-45.

Feinberg, T.E. 1997. The irreducible perspectives of consciousness. *Seminars in Neurology* 17:85-93.

Fodor, J.A. 1981. The mind-body problem. *Scientific American* 244:114-25.
An overview: behaviorism, identity theory, functionalism, etc.

Foss, J.E. 1987. Is the mind-body problem empirical? Canadian Journal of Philosophy 17:505-32.

Yes it is. Empirical evidence bears on materialism, property dualism, emergentism, functionalism, interactive dualism, idealism, etc.

Gomes, G. 1995. Self-awareness and the mind-brain problem. Philosophical Psychology 8:155-65.

Gunderson, K. 1970. Asymmetries and mind-body perplexities. Minnesota Studies in the Philosophy of Science 4:273-309.

The core of the mind-body problem is the first/third-person asymmetry. It's like a periscope trying to place itself between its crosshairs. But this doesn't imply any strong ontological consequences.

Honderich, T. 1989. Mind and Brain. Oxford University Press.

Honderich, T. 1995. Consciousness, neural functionalism, and real subjectivity. American Philosophical Quarterly 32:369-381.

Against "neural functionalism", and on how Searle's view reduces to either neural functionalism or property dualism.

Howard, D.J. 1986. The new mentalism. International Philosophical Quarterly 26:353-7.

Hutto, D.D. 1998. An ideal solution to the problems of consciousness. Journal of Consciousness Studies 5:328-43.

Kim, J. 1997. The mind-body problem: Taking stock after forty years. Philosophical Perspectives 11:185-207.

Kneale, M. 1950. What is the mind-body problem? Proceedings of the Aristotelian Society 50:105-22.

Kohler, W. 1960. The mind-body problem. In (S. Hook, ed) Dimensions of Mind. New York University Press.

Levin, M. 1979. Metaphysics and the Mind-Body Problem. Oxford University Press.

Lockwood, M. 1989. Mind, Brain, and the Quantum. Oxford University Press.
On the mind-body problem and physical theory. Against reductive physicalism; instead, experience is the intrinsic nature of the physical. Discusses the interpretation of quantum mechanics and much else. A very enjoyable book.

Lowe, E.J. 1996. Subjects of Experience. Cambridge University Press.

Lund, D.H. 1994. Perception, Mind, and Personal Identity: A Critique of Materialism. University Press of America.

Margolis, J. 1974. Reductionism and ontological aspects of consciousness. Journal for the Theory of Social Behavior 4:3-16.

Matson, W.I. 1966. Why isn't the mind-body problem ancient? In (P. Feyerabend & G. Maxwell, eds) Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl. University of Minnesota Press.

Matson, W.I. 1976. Sentience. University of California Press.

McMullen, T. 1997. Sperry on consciousness as an emergent causal agent. Australian Journal of Psychology 49:152-155.

Nagel, T. 1993. What is the mind-body problem? In Experimental and Theoretical Studies of Consciousness (Ciba Foundation Symposium 174). Wiley.
On ways in which we might locate consciousness within the natural world via

scientific study. Perhaps we need an wider conception of objective reality.

Nagel, T. 1994. Consciousness and objective reality. In (R. Warner & T. Szubka, eds) The Mind-Body Problem: A Guide to the Current Debate. Blackwell.

Nagel, T. 2001. The psychophysical nexus. In (P. Boghossian & C. Peacocke, eds) New Essays on the A Priori. Oxford University Press.

O'Shaughnessy, B. 1994. The mind-body problem. In (R. Warner & T. Szubka, eds) The Mind-Body Problem: A Guide to the Current Debate. Blackwell.

Pratt, J.B. 1936. The present status of the mind-body problem. *Philosophical Review* 65:144-56.

Reber, A. 1997. Caterpillars and consciousness. *Philosophical Psychology* 10:437-49.

Ripley, C. 1984. Sperry's concept of consciousness. *Inquiry* 27:399-423.
An in-depth analysis of Sperry's views on consciousness. Sperry is not a dualist; he simply believes in "structural causation" based on emergent properties. Thorough and interesting.

Robinson, W.S. 1988. Brains and People: An Essay on Mentality and its Causal Conditions. Temple University Press.

Rosenthal, D.M. (ed) 1971. Materialism and the Mind-Body Problem. Prentice-Hall.
A collection of essays from the 1960s on the identity theory, functionalism, eliminative materialism.

Senchuk, D.M. 1991. Consciousness naturalized: Supervenience without physical determinism. *American Philosophical Quarterly* 28:37-47.

Sellars, W. 1953. A semantical solution of the mind-body problem. *Methodos* 5:45-84. Reprinted in Pure Pragmatics and Possible Worlds. Ridgview, 1980.

Sellars, W. 1971. The double knowledge approach to the mind-body problem. *New Scholasticism* 45:269-89.

Silberstein, M. 1998. Emergence and the mind-body problem. *Journal of Consciousness Studies* 5:464-82.

Sperry, R.W. 1969. A modified concept of consciousness. *Psychological Review* 76:532-36.
Consciousness is an emergent property of brain dynamics that itself governs low-level flow of excitation. Midway between mentalism and materialism.

Sperry, R.W. 1980. Mind-brain interaction: Mentalism yes, dualism no. *Neuroscience* 5:195-206.
A summary of the position whereupon mental properties are emergent and have independent causal powers. With a contrast to Popper and Eccles' dualism.

Sperry, R.W. 1992. Turnabout on consciousness: A mentalist view. *Journal of Mind and Behavior* 13:259-80.
An account of the "new mentalist paradigm". Clarifies earlier work, comments on others' interpretations. The view is monist and functionalist, but consciousness is a distinct emergent quality with a "downward" causal role.

Strawson, G. 1994. The experiential and the non-experiential. In (R. Warner & T. Szubka, eds) The Mind-Body Problem: A Guide to the Current Debate. Blackwell.

Velmans, M. 1990. Consciousness, brain, and the physical world. *Philosophical Psychology* 3:77-99.

Wagner, S.J. 1994. Supervenience, recognition, and consciousness. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Warner, R. & Szubka, T. 1994. *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

A collection of 27 (mostly original) papers on the mind-body problem.

Weintraub, R. 1999. The spatiality of the mental and the mind-body problem. *Synthese* 117:409-17.

Wilson, D.L. 1976. On the nature of consciousness and of physical reality. *Perspectives in Biology and Medicine* 19:568-581.

Wisdom, J. 1957. Some main mind-body problems. *Proceedings of the Aristotelian Society* 60:187-210.

1.4 Specific Views on Consciousness [see also 1.2, 1.3, 1.5c]

1.4a Higher-Order Thought Approaches (Rosenthal, etc) [see also 6.2i]

Aquila, R. 1990. Consciousness as higher-order thoughts: Two objections. *American Philosophical Quarterly* 27:81-87.

Higher-order thought theories have two unacceptable consequences: one can notice one's hearing a sound without noticing one's consciousness of the sound; and one can unconsciously perceive one's surroundings as gloomy.

Byrne, A. 1997. Some like it HOT: consciousness and higher-order thoughts. *Philosophical Studies* 2:103-29.

Carruthers, P. 1989. Brute experience. *Journal of Philosophy* 258-69.

Argues for a distinction between conscious and non-conscious experiences, depending on whether one is conscious of the experience. Animal experiences are of the second kind, and therefore are not morally significant.

Carruthers, P. 1992. Consciousness and concepts. *Aristotelian Society Supplement* 66:41-59.

Advocates the "reflexive thinking" account of consciousness over Kirk's "presence" account. Availability for reflexive thinking is naturally necessary and sufficient for qualia. Interesting paper.

Carruthers, P. 1996. *Language, Thought, and Consciousness*. Cambridge University Press.

Carruthers, P. 1997. Fragmentary versus reflexive consciousness. *Mind and Language* 12:181-95.

Carruthers, P. 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.

Dretske, F. 1993. Conscious experience. *Mind* 102:263-283.

Against higher-order thought accounts: one can have a conscious experience without being aware that one is having it. With remarks on thing-awareness vs. fact-awareness and on "inner-sense" accounts.

Dretske, F. 1995. Are experiences conscious? In *Naturalizing the Mind*. MIT Press.

We're not conscious *of* our experience in general, but conscious *with* it.

Criticizes HOP theories (not conceptualized enough) and HOT theories (rules out animals; there's more in experience than thought).

Francescotti, R.M. 1995. Higher-order thoughts and conscious experience. *Philosophical Psychology*.

Argues that a higher-order thought is insufficient for consciousness, even with Rosenthal's constraint. A causal constraint is required, but the only strong enough such constraint doesn't work.

Gennaro, R.J. 1993. Brute experience and the higher-order thought theory of consciousness. *Philosophical Papers* 22:51-69.

Carruthers 1989 misanalyzes higher-order thought theory. There's no need for conscious HOTs, and not too much conceptual sophistication is required, so animals might have HOTs and therefore conscious pains.

Gennaro, R.J. 1996. *_Consciousness and Self-consciousness: A Defense of the Higher-Order Thought Theory of Consciousness_*. John Benjamins.

Guzeldere, G. 1996. Consciousness and the introspective link principle. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Guzeldere, G. 1995. Is consciousness the perception of what passes in one's own mind? In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

A critique of higher-order-perception theories of consciousness. They're either committed to a "representational divide" fallacy or collapse into higher-order-thought or first-order theories.

Jamieson, D. & Bekoff, M. 1992. Carruthers on nonconscious experience. *Analysis* 52:23-28.

Various points against Carruthers 1989. His examples of nonconscious experience are likely conscious, and the higher-order account is circular.

Jacob, P. 1996. State consciousness revisited. *Acta Analytica* 11:29-54.

Kobes, B.W. 1995. Telic higher-order thoughts and Moore's paradox. *Philosophical Perspectives* 9:291-312.

Levine, J. 1997. Are qualia just representations? (Critical notice of Tye.) *Mind and Language* 12:101-13.

Lycan, W.G. 1995. Consciousness as internal monitoring, I. *Philosophical Perspectives* 9:1-14.

Argues for a Lockean quasi-perceptual view of consciousness as internal monitoring via second-order states. Contra objections, e.g. Rey's point that it makes consciousness too prevalent -- consciousness isn't an on-off affair.

Lycan, W. 2001. A simple argument for a higher-order representation theory of consciousness. *Analysis* 61:3-4.

Mellor, D.H. 1978. Conscious belief. *Proceedings of the Aristotelian Society* 78:87-101.

Conscious belief (or assent) is believing that one believes. Addresses various objections, from self-deception and from consciousness of assent. Communication needs conscious belief, not just belief.

Mellor, D.H. 1980. Consciousness and degrees of belief. In (D.H. Mellor, ed) *_Prospects for Pragmatism_*. Cambridge University Press.

Natsoulas, T. 1992. Appendage theory -- pro and con. *Journal of Mind and Behavior* 13:371-96.

On various pros and cons of HOT theories, to do with reflexivity, objects of HOTs, introspection, and so on. With comparisons to "intrinsic" theories.

Natsoulas, T. 1992. Are all instances of phenomenal experience conscious in the sense of their being objects of inner (second-order) consciousness? *American Journal of Psychology* 105:605-12.

Natsoulas, T. 1993. What is wrong with the appendage theory of consciousness? *Philosophical Psychology* 6:137-54.

On three theories of our direct awareness of conscious states: mental-eye theories, self-intimational theories, and appendage theory. Appendage theory (i.e. HOT theory) is promising, but how does an HOT determine its object?

Natsoulas, T. 1993. The importance of being conscious. *Journal of Mind and Behavior* 14:317-40.

On the differences between first-order and second-order consciousness. Second-order consciousness is essential for communication and locomotion. With remarks on "nonconscious consciousness".

Nelkin, N. 1989. Unconscious sensations. *Philosophical Psychology* 2:129-41. Separates CN (phenomenological) from C1 (info-processing) and C2 (higher order beliefs). CN is a subset of CS (image-representation state). We are always C2 of CN states, but not of other CS states: unconscious sensations!

Nelkin, N. 1995. The dissociation of phenomenal states from apperception. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Argues that we need not be apperceptively aware of phenomenal states. Introspection leaves the matter open, but some empirical results (e.g. hue blindsight) and theoretical arguments support dissociability.

Ridge, M. 2001. Taking solipsism seriously: Nonhuman animals and meta-cognitive theories of consciousness. *Philosophical Studies* 103:315-340.

Rosenthal, D.M. 1986. Two concepts of consciousness. *Philosophical Studies* 49:329-59.

Consciousness should be construed neither as sensation nor intentionality, but as the existence of higher-order thoughts.

Rosenthal, D.M. 1997. A theory of consciousness. In (N. Block, O. Flanagan, and G. Guzeldere, eds) *_The Nature of Consciousness_*. MIT Press.

A conscious mental state is a state that is the subject of a higher-order thought. Consciousness is not essential to mentality, should be separated from sensory quality, and is not an intrinsic property of conscious states.

Rosenthal, D.M. 1990. Why are verbally expressed thoughts conscious? *Bielefeld Report*.

Because verbally expressing and reporting are easily and immediately connected for 1st-order thoughts. But not for 2nd-order thoughts. Hmmm.

Rosenthal, D.M. 1993. Thinking that one thinks. In (M. Davies and G. Humphreys, eds) *_Consciousness: Psychological and Philosophical Essays_*. Blackwell.

Conscious states are states that are the contents of higher-order thoughts. Express/report distinction: we report them, and express the HOT (which may be unconscious). Defense against dispositional and collapsing objections.

Rosenthal, D.M. 1993. Explaining consciousness. Manuscript.

Distinguishes the sense in which we are aware of conscious states; argues for the separation of consciousness and sensation; and outlines how higher-order thoughts might explain the what-it's-like of conscious states.

Rosenthal, D.M. 1993. Higher-order thoughts and the appendage theory of consciousness. *Philosophical Psychology* 6:155-66.

In response to Natsoulas, HOT theory needn't answer the general question of how intentional states determine their objects. With remarks on the

other alternatives and the dangers of self-intimation.

Rosenthal, D.M. 1995. Moore's paradox and consciousness. *Philosophical Perspectives* 9:313-33.

Rosenthal, D.M. 1997. Apperception, sensation, and dissociability. *Mind and Language* 2:206-23.

Rosenthal, D.M. 1998. Consciousness and metacognition. In (D. Sperber, ed) *_Metarepresentation_*. Oxford University Press.

Rosenthal, D.M. 2000. Consciousness, interpretation, and consciousness. *Protosociology* 14.

Rowlands, M. 2001. Consciousness and higher-order thoughts. *Mind and Language* 16:290-310.

Seager, W.E. 1994. Dretske on HOT theories of consciousness. *Analysis* 54:270-76.

Seager, W.E. 1999. HOT Theory: The mentalistic reduction of consciousness. In *_Theories of Consciousness: An Introduction and Assessment_*. Routledge.

Stamenov, M.I. 1997. Grammar, meaning, and consciousness: What sentence structure can tell us about the structure of consciousness. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

Stoerig, P. 1997. Phenomenal vision and apperception: Evidence from blindsight. *Mind and Language* 2:224-37.

Stone, J. 2001. What is it like to have an unconscious mental state? *Philosophical Studies* 104:179-202.

Thomasson, A. 2000. After Brentano: A one-level theory of consciousness. *European Journal of Philosophy* 8:190-210.

1.4b Dennett on Consciousness [see also 1.7c]

Akins, K. 1996. Lost the plot? Reconstructing Dennett's multiple drafts theory of consciousness. *Mind and Language* 11:1-43.

Akins, K, & Winger, S. 1996. Ships in the night: Churchland and Ramachandran on Dennett's theory of consciousness. In (K. Akins, ed) *_Perception_*. Oxford University Press.

Arbib, M.A. 1972. Consciousness: The secondary role of language. *Journal of Philosophy* 69.

Baker, L.R. 1995. Content meets consciousness. *Philosophical Topics* 22:1-22.

Block, N. 1995. What is Dennett's theory a theory of? *Philosophical Topics* 22:23-40.

Bricke, J. 1984. Dennett's eliminative arguments. *Philosophical Studies* 45:413-29.

Criticizing Dennett's accounts of pains, dreams, and images: in no case do his arguments earn their eliminative conclusions.

Bricke, J. 1985. Consciousness and Dennett's intentionalist net. *Philosophical Studies* 48:249-56.

Reportability is no good for capturing consciousness: it completely leaves out the qualitative content of conscious states.

Churchland, P.S. & Ramachandran, V.S. 1993. Filling in: Why Dennett is wrong. In (B. Dahlbom, ed) Dennett and His Critics. Blackwell.
Argues that Dennett's account of the blindspot and scotomas are wrong.
Neurophysiological data suggests that blind areas are represented explicitly; psychological data shows that it's not just "more of the same".

Clark, S.R.L. 1993. Minds, memes, and rhetoric. *Inquiry* 36:3-16.

Dennett, D.C. 1968. Content and Consciousness. Routledge.

Dennett, D.C. 1978. Reply to Arbib and Gunderson. In Brainstorms. MIT Press.
On various notions of awareness: contents of the speech center, contents directing behavior, and contents of attention. We have privileged access to one sort, but it is a different sort that plays the main role in control.

Dennett, D.C. 1978. Toward a cognitive theory of consciousness. *Minnesota Studies in the Philosophy of Science*, Vol. 9. Reprinted in Brainstorms (MIT Press, 1978).
Conscious contents are contents of a buffer memory to which a public report module has access. We only have conscious access to propositional judgments, not to underlying processes. With a cute functional diagram.

Dennett, D.C. 1979. On the absence of phenomenology. In (D. Gustafson & B. Tapscott, eds) Body, Mind, and Method. Kluwer.
There is no real phenomenology. There are only *judgments* about phenomenology, and nothing more is going on. We don't have privileged access to anything, except perhaps certain propositional episodes.

Dennett, D.C. 1982. How to study human consciousness empirically, or, nothing comes to mind. *Synthese* 53:159-80.
We can study consciousness by the method of heterophenomenology: studying the things we say about conscious states, which we can interpret as we interpret texts. Autophenomenology gives nothing extra. With comments by Rorty.

Dennett, D.C. 1988. The evolution of consciousness. Manuscript.
Consciousness is a virtual machine which evolved.

Dennett, D.C. 1991. Consciousness Explained. Little-Brown.
Argues against the "Cartesian Theatre", advocating a "multiple drafts" model of consciousness. Presents a detailed model of processes underlying verbal report, and argues that there is nothing else (e.g. qualia) to explain.

Dennett, D.C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15:183-201.
Using temporal anomalies in consciousness to support a "Multiple Drafts" theory of consciousness rather than a "Cartesian Theater". Contents of consciousness are wholly determined by effects on action/memory.

Dennett, D.C. 1993. *Precis of Consciousness Explained*. *Philosophy and Phenomenological Research* 53:889-931.
A discussion of Consciousness Explained, with comments by Tye, Jackson, Shoemaker, and Rosenthal, and a reply by Dennett.

Dennett, D.C. 1993. Living on the edge. *Inquiry* 36:135-59.
A reply to Clark, Fellows & O'Hear, Foster, Lockwood, Seager, Siewert, and Sprigge.

Dennett, D.C. 1993. Caveat emptor. *Consciousness and Cognition* 2:48-57.
A reply to Baars & McGovern, Mangan, Toribio.

Dennett, D.C. 1995. Is perception the "leading edge" of memory. In (A.

Spafadora, ed) Memory and Oblivion.

There is no "leading edge" of consciousness, separating perception and memory. With an analysis of metacontrast cases, etc.

Dennett, D.C. 1995. Get real. *Philosophical Topics* 22:505

Dennett, D.C. 1996. Seeing is believing -- or is it? In (K. Akins, ed) Perception. Oxford University Press.

Dennett, D. 2001. Are we explaining consciousness yet? *Cognition* 79:221-37.

Densmore, S. & Dennett, D.C. 1999. The virtues of virtual machines. *Philosophy and Phenomenological Research* 59:747-61.

Dretske, F. 1995. Differences that make no difference. *Philosophical Topics* 22:41-57.

Criticizes Dennett's first-person operationalism as Cartesian. There can be awareness without judgment -- e.g. non-epistemic perception. This comes from information or "micro-judgments", and is not conceptual.

Fellows, R. & O'Hear, A. 1993. Consciousness avoided. *Inquiry* 36: 73-91.

Foster, J. 1993. Dennett's rejection of dualism. *Inquiry* 36:17-31.

Gunderson, K. 1972. Content and Consciousness and the mind-body problem. *Journal of Philosophy* 69.

Hutto, D. 1995. Consciousness demystified: A Wittgensteinian critique of Dennett. *Monist* 78:464-79.

Jackson, F. 1993. Appendix A (for philosophers). *Philosophy and Phenomenological Research* 53:897-901.

Presses Dennett on the "truth-maker" question for materialists: what sort of physical fact makes it true that people are conscious?

Johnsen, B. 1997. Dennett on qualia and consciousness: A critique. *Canadian Journal of Philosophy* 27:47-82.

Kirk, R. 1993. "The best set of tools"? Dennett's metaphors and the mind-body problem. *Philosophical Quarterly* 43:335-43.

Joycean machines and multiple drafts turn out to shed no light on the question of what features make a conscious system conscious.

Lockwood, M. 1993. Dennett's mind. *Inquiry*.

Argues for a suitably sophisticated Cartesian Theatre, and against the identification of phenomenology with judgments.

Mangan, B. 1993. Dennett, consciousness, and the sorrows of functionalism. *Consciousness and Cognition* 2:1-17.

Marbach, E. 1988. How to study consciousness phenomenologically or quite a lot comes to mind. *Journal of the British Society for Phenomenology*, 19:252-268.

Marbach, E. 1994. Troubles with heterophenomenology. In (R. Casati, B. Smith, & S. White, eds) Philosophy and the Cognitive Sciences. Holder-Pichler-Tempsky.

McCauley, R.N. 1993. Why the blind can't lead the blind: Dennett on the blind spot, blindsight, and sensory qualia. *Consciousness and Cognition* 2:155-64.

Brings empirical evidence to bear against Dennett's "filling-in" account of the blindspot, and argues that blindsight and the blindspot aren't analogous.

McGinn, C. 1995. Consciousness evaded: Comments on Dennett. *Philosophical Perspectives* 9:241-49.

Nikolinakos, D. 2000. Dennett on qualia: The case of pain, smell and taste. *Philosophical Psychology* 13:505-522.

Puccetti, R. 1993. Dennett on the split-brain. *Psychology* 4(52).

Radner, D. 1994. Heterophenomenology: Learning about the birds and the bees. *Journal of Philosophy* 91:389-403.

Rey, G. 1995. Dennett's unrealistic psychology. *Philosophical Topics* 22:259-89.

Robinson, W.S. 1972. Dennett's analysis of awareness. *Philosophical Studies* 23:147-52.

Robinson, W.S. 1994. Orwell, Stalin, and determinate qualia. *Pacific Philosophical Quarterly* 75:151-64.

Dennett's Orwell/Stalin argument doesn't establish its conclusion, as "brain smear" is quite compatible with determinate qualia.

Rockwell, T. 1996. Awareness, mental phenomena, and consciousness: A synthesis of Dennett and Rosenthal. *Journal of Consciousness Studies* 3:463-76.

Rorty, R. 1972. Dennett on awareness. *Philosophical Studies* 23:153-62.

Rosenthal, D.M. 1993. Multiple drafts and higher-order thoughts. *Philosophy and Phenomenological Research* 53:911-18.

Rosenthal, D.M. 1994. First-person operationalism and mental taxonomy. *Philosophical Topics* 22:319-349.

Rosenthal, D.M. 1995. Multiple drafts and the facts of the matter. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

We can subtract first-person operationalism from Dennett's multiple-drafts account, giving a higher-order thought theory.

Rosenthal, D. 2000. Content, interpretation, and consciousness. *Protosociology* 14:67-84.

Seager, W.E. 1993. Verification, skepticism, and consciousness. *Inquiry*. An elucidation of Dennett's fundamental eliminativism about phenomenology, resting on verificationist arguments. Like many sceptical arguments, it ends up too powerful to be convincing.

Seager, W.E. 1999. Dennett, part I and II. In *Theories of Consciousness: An Introduction and Assessment*. Routledge.

Shoemaker, S. 1993. Lovely and suspect ideas. *Philosophy and Phenomenological Research* 53:903-908.

Siewert, C. 1993. What Dennett can't imagine and why. *Inquiry*. Argues that zombies are conceivable, via partial zombiehood in blindsight patients who respond unprompted. Dennett's arguments rely on a question-begging third-person absolutism.

Sprigge, T.L.S. 1993. Is Dennett a disillusioned zimbo? *Inquiry* 36:33-57.

Toribio, J. 1993. Why there still has to be a theory of consciousness. *Consciousness and Cognition* 2:28-47.

Criticizes behavioral, localist, and "intransitive" approaches to consciousness, and recommends a "transitive" metacognitive approach. But criticizes Dennett for not explaining subjective experience.

Tye, M. 1993. Reflections on Dennett and consciousness. *Philosophy and*

Phenomenological Research 53:891-6.

Argues that Dennett's verificationism begs the question, and that "seeming" cannot be identified with believing or judging.

Van Gulick, R. 1995. Dennett, drafts, and phenomenal realism. Philosophical Topics 22:443-55.

Wuketits, F. 1994. Consciousness explained -- or explained away? Acta Analytica 9:55-64.

1.4c Functionalism [see also 1.4a, 1.4b, 1.5c, 1.8, 3.4]

Levin, J. 1991. Analytic functionalism and the reduction of phenomenal states. Philosophical Studies 61:211-38.

Contra Kripkean arguments, a good enough functional theory may help close the conceivability/explanatory gap between the physical and qualia. Contra Nagel/Jackson, such a theory could provide us with recognitional abilities.

Mangan, B. 1998. Against functionalism: Consciousness as an information-bearing medium. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness II. MIT Press.

Marcel, A. 1988. Phenomenal experience and functionalism. In (A. Marcel & E. Bisiach, eds) Consciousness in Contemporary Science. Oxford University Press.

Myin, E. 1998. Holism, functionalism and visual awareness. Communication and Cognition, 31:3-19.

Perlis, D. 1995. Consciousness and complexity: The cognitive quest. Annals of Mathematics and Artificial Intelligence 14:309-21.

Shoemaker, S. 1993. Functionalism and consciousness. In Experimental and Theoretical Studies of Consciousness (Ciba Foundation Symposium 174). Wiley. Argues that introspective access is essential to many sorts of mental state, due to constitutive rationality requirements. Against a perceptual model of introspection; introspecting and introspected states are closer than that.

Schweizer, P. 1996. Physicalism, functionalism, and conscious thought. Minds and Machines 6:61-87.

van Gulick, R. 1988. A functionalist plea for self-consciousness. Philosophical Review 97:149-88.

How functionalism can handle consciousness: Self-consciousness is the possession of reflexive metapsychological information. This helps understand learning, representation, and belief. Phenomenal experience is still tricky.

1.4d Eliminativism [see also 1.7c, 2.1c, 3.5c]

Allport, A. 1988. What concept of consciousness? In (A. Marcel & E. Bisiach, eds) Consciousness in Contemporary Science. Oxford University Press.

Churchland, P.S. 1983. Consciousness: the transmutation of a concept. Pacific Philosophical Quarterly 64:80-95.

Experimental evidence against consciousness/introspection/transparency.

Dennett, D.C. 1976. Are dreams experiences? Philosophical Review 73:151-71. Reprinted in Brainstorms (MIT Press, 1978).

Argues that dreams might not be experienced, but rather be stored directly into memory (the "cassette-tape" theory of dreaming).

Dennett, D.C. 1979. The onus re experiences: A reply to Emmett. *Philosophical Studies* 35, 315-18.

Emmett, K. 1978. Oneiric experiences. *Philosophical Studies* 34:445-50.

Nikolinakos, D. 1994. General anesthesia, consciousness, and the skeptical challenge. *Journal of Philosophy* 2:88-104.

Consciousness is an indispensable concept in anesthesiology, and therefore (contra Churchland and Wilkes) is a scientifically legitimate kind. With empirical details and anesthesiological theory on levels of consciousness.

Rey, G. 1982. A reason for doubting the existence of consciousness. In (R. Davidson, S. Schwartz, & D. Shapiro, eds) *_Consciousness and Self-Regulation_*, Vol 3. Plenum Press.

One could make a machine, duplicating the usual abilities that go along with consciousness, but surely it wouldn't be conscious. So what are the conditions for consciousness? Maybe there are none.

Rey, G. 1986. A question about consciousness. In (H. Otto & J. Tuedio, eds) *_Perspectives on Mind_*. Kluwer.

A rerun of Rey 1982: An unconscious machine could duplicate all the obvious criteria for consciousness, so maybe even we aren't conscious. With remarks on the relation between our belief in consciousness and consciousness itself.

Rey, G. 1995. Toward a projectivist account of conscious experience. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

We "project" consciousness into ourselves and others. There are no explanation-transcendent phenomena for which there is non-question-begging evidence. With remarks on self-attribution and Wittgenstein.

Smith, D.W. 1986. In (H. Otto & J. Tuedio, eds) *_Perspectives on Mind_*. Kluwer.

Commentary on Rey 1986: we are directly aware of our consciousness. It's not a theoretical entity, but rather something to be explained.

Tienison, J.L. 1987. Brains are not conscious. *Philosophical Papers* 16:187-93.

A skeptical argument: single neurons are not conscious, and adding a neuron won't produce consciousness, so finite brains are not conscious.

Wilkes, K.V. 1984. Is consciousness important? *British Journal for the Philosophy of Science* 35:223-43.

No, and it's not very coherent either. It divides into awakesness, sensation, sensory experience, and propositional attitudes. Also a history of the term.

Wilkes, K.V. 1988. Yishi, Duh, Um and consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Wilkes, K.V. 1995. Losing consciousness. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Consciousness is not a tenable notion in either commonsense or scientific psychology; we should return instead to the "psuche".

Williams, D.C. 1934. Scientific method and the existence of consciousness. *Psychological Review* 41:461-79.

Williams, D.C. 1959. Mind as a matter of fact. *Review of Metaphysics* 13:205-25.

1.4e Epiphenomenalism [see also 1.3a, 1.3f, 1.6e, 3.6]

Bieri, P. 1992. Trying out epiphenomenalism. *Erkenntnis* 36:283-309.

Birnbacher, D. 1988. Epiphenomenalism as a solution to the ontological mind-body problem. *Ratio* 1:17-32.

Chalmers, D.J. 1996. The paradox of phenomenal judgment. In *The Conscious Mind*. Oxford University Press.

Considers major arguments against the causal or explanatory irrelevance of consciousness -- arguments from self-knowledge, memory, reference, etc -- and argues that none pose fatal flaws.

Creel, R. 1980. Radical epiphenomenalism: B.F. Skinner's account of private events. *Behaviorism* 8:31-53.

Dennett, D.C. 1991. "Epiphenomenal" qualia? In *Consciousness Explained*, pp. 398-406. Little-Brown.

Discusses two senses of "epiphenomenalism" -- "Huxley's" and "philosophical" varieties -- and argues that the philosophical sort is crazy. (N.B. Huxley actually subscribed to the "philosophical" variety.)

Double, R. 1979. Taylor's refutation of epiphenomenalism. *Journal of Critical Analysis* 8:23-28.

Flanagan, O. & Polger, G. 1998. Consciousness, adaptation, and epiphenomenalism. In (G. Mulhauser, ed) *Evolving Consciousness*.

Hodges, M. 1979. Meaning and the impotence hypothesis. *Review of Metaphysics* 32:515-29.

Horowitz, A. 1999. Is there a problem in physicalist epiphenomenalism? *Philosophy and Phenomenological Research* 59:421-34.

Huxley, T. 1874. On the hypothesis that animals are automata. *Fortnightly Review* 95:555-80. Reprinted in *Collected Essays*. London, 1893.

Hyslop, A. 1998. Methodological epiphenomenalism. *Australasian Journal of Philosophy* 78:61-70.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Uses the Mary thought-experiment to argue that qualia are epiphenomenal, and argues that epiphenomenalism is a tenable doctrine.

Kraemer, E.R. 1980. Imitation-man and the 'new' epiphenomenalism. *Canadian Journal of Philosophy* 10:479-487.

If Campbell's imitation man is possible, then the causal relation between the physical and phenomenal is unreliable.

Lachs, J. 1963. Epiphenomenalism and the notion of cause. *Journal of Philosophy* 60:141-45.

Lachs, J. 1963. The impotent mind. *Review of Metaphysics* 17:187-99.

Lachs, J. 1967. Angel, animal, machine: Models for man. *Southern Journal of Philosophy* 5:221-27.

Long, W. 1953. Comments on the alleged proof of epiphenomenalism. *British Journal for the Philosophy of Science* 3:355-58.

Pauen, M. 2000. Painless pain: Property dualism and the causal role of phenomenal consciousness. *American Philosophical Quarterly* 37:51-64.

Pecnjak, D. 1989. Epiphenomenalism and machines: A discussion of van Rooijen's critique of Popper. *British Journal for the Philosophy of Science* 40:404-8.

Popper, K.R. 1977. Some remarks on panpsychism and epiphenomenalism. *Dialectica* 31:177-86.

Puccetti, R. 1974. Physicalism and the evolution of consciousness. *Canadian Journal of Philosophy Supplement* 1:171-83.

Robinson, D. 1993. Epiphenomenalism, laws, and properties. *Philosophical Studies* 69:1-34.

Argues that phenomenal information implies epiphenomenalism, even at the intra-psychic level. With remarks on ineffability and on whether properties should be individuated by nomic role or by essence.

Robinson, W.S. 1982. Causation, sensation, and knowledge. *Mind* 91:524-40.

Argues that we can have non-inferential knowledge of sensations even if they don't make a causal difference to our beliefs, so epiphenomenalism is OK. All theories have a similar problems.

Rudd, A. 2000. Phenomenal judgment and mental causation. *Journal of Consciousness Studies* 7:53-69.

van Rooijen, J. 1987. Interactionism and evolution: A critique of Popper. *British Journal for the Philosophy of Science* 38:87-92.

Wasserman, G.D. 1982. Materialism and mentality. *Review of Metaphysics* 35:715-30.

Wassermann, G. 1979. Reply to Popper's attack on epiphenomenalism. *Mind* 88:572-75.

Watkins, M. 1989. The knowledge argument against the knowledge argument. *Analysis*, 49:158-60.

Epiphenomenalism => qualia don't cause beliefs => we don't know about qualia.

Wisdom, J.O. 1954. Is epiphenomenalism refutable? *Proceedings of the 2nd International Congress of the International Union for the Philosophy of Science* 5:73-78.

Woodhouse, M. 1974. A new epiphenomenalism? *Australasian Journal of Philosophy* 52:163-69.

1.4f Interactionism [see also 1.3f, 3.3d]

Beloff, J. 1994. Minds and machines: A radical dualist perspective. *Journal of Consciousness Studies* 1:32-37.

Buncombe, M. 1995. *_The Substance of Consciousness: An Argument for Interactionism_*. Avebury.

Elitzur, A.C. 1989. Consciousness and the incompleteness of the physical explanation of behavior. *Journal of Mind and Behavior* 10:1-20.

Argues from the fact that we talk about consciousness to the conclusion that consciousness plays an active role, so physical laws must be incomplete.

Elitzur, A.C. 1990. Neither idealism nor materialism: A reply to Snyder. *Journal of Mind and Behavior*.

Elitzur, A.C. 1995. Consciousness can no longer be ignored. *Journal of Consciousness Studies* 2:353-58.

Foster, J. 1991. *_The Immaterial Self: A Defense of the Cartesian Dualist Conception of Mind_*. Routledge.

Hodgson, D. 1991. *_The Mind Matters: Consciousness and Choice in a Quantum*

World_. Oxford University Press.

Jackson, F. 1980. Interactionism revived? *Philosophy of Social Science* 10:316-23.

Libet, B. 1994. A testable theory of mind-brain interaction. *Journal of Consciousness Studies* 1:119-26.

Lindahl, B.I.B. & Arhem, P. 1996. Mind as a force field: Comments on a new interactionistic hypothesis. *Journal of Theoretical Biology* 171:111-22.

Popper, K. & Eccles, J. 1977. *The Self and Its Brain: An Argument for Interactionism*. Springer.

Popper, K. 1994. *Knowledge and the Body-Mind Problem: In Defence of Interaction*. Routledge.

Roelofs, H.D. 1955. A case for dualism and interactionism. *Philosophy and Phenomenological Research* 15:451-76.

Snyder, D. 1990. On Elitzur's discussion of the impact of consciousness on the physical world. *Journal of Mind and Behavior*.
Argues with Elitzur on quantum mechanics and consciousness. With response.

Warner, R. 1996. Facing ourselves: Incorrigeability and the mind-body problem. *Journal of Consciousness Studies* 3:217-30.

Wilson, D.L. 1999. Mind-brain interactionism and the violation of physical laws. *Journal of Consciousness Studies*.

1.4g Panpsychism [see also 1.4h]

Bjelland, A.G. 1982. Popper's critique of panpsychism and process proto-mentalism. *Modern Schoolman* 59:233-43.

Butler, C. 1978. Panpsychism: A restatement of the genetic argument. *Idealist Studies* 8:33-39.

Chalmers, D.J. 1996. Is experience ubiquitous? In *The Conscious Mind*. Oxford University Press.

There are no strong arguments against panpsychism, and good reason to take it seriously. Extrapolating the processing properties crucial for standard complex experience suggests that simple process may yield simple experience.

de Quincey, C. 1994. Consciousness all the way down? *Journal of Consciousness Studies* 1:217-29.

An analysis of a debate between Griffin and McGinn on panexperientialism, arguing for new forms of understanding.

Edwards, P. 1967. Panpsychism. In (P. Edwards, ed) *The Encyclopedia of Philosophy*, volume 5. Macmillan.

An excellent review article on panpsychism; highly recommended.

Farleigh, P. 1998. Whitehead's even more dangerous idea. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Ford, M.P. 1981. William James: Panpsychist and metaphysical realist. *Transactions of the Peirce Society* 17:158-70.

Griffin, D.R. 1997. Panexperiential physicalism and the mind-body problem. *Journal of Consciousness Studies* 4:248-68.

An interesting paper arguing for an experiential aspect in all matter,

explicating a Whiteheadian position.

Griffin, D.R. 1998. *_Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem_*. University of California Press.

Griffin, D.R. 1998. Pantemporalism and panexperientialism. In (P. Harris, ed) *_The Textures of Time_*. University of Michigan Press.

Hartshorne, C. 1978. Panpsychism: Mind as sole reality. *Ultim Real Mean* 1:115-29.

Hut, P. & Shepard, R. 1996. Turning the "hard problem" upside-down and sideways. *Journal of Consciousness Studies* 3:313-29.

Argues for a new fundamental feature ("X") which stands to consciousness as time stands to motion, thus making consciousness possible and ubiquitous.

Nagel, T. 1979. Panpsychism. In *_Mortal Questions_*. Cambridge University Press.

Material composition, nonreductionism, realism, non-emergence => panpsychism.

Popper, K.R. 1977. Some remarks on panpsychism and epiphenomenalism. *Dialectica* 31:177-86.

Rensch, B. 1977. Argument for panpsychist identism. In (J. Cobb & D. Griffin, eds) *_Mind in Nature_*. University Press of America.

Robinson, E.A. 1949. Animism as a world hypothesis. *Philosophical Review* 58:53-63.

Rosenberg, G.H. 1996. Rethinking nature: A hard problem within the hard problem. *Journal of Consciousness Studies* 3:76-88.

On why consciousness extends beyond the cognitive. Argues that fundamental laws for consciousness must connect at a basic level, and argues that panpsychism is not as implausible as often thought.

Seager, W. 1995. Consciousness, information, and panpsychism. *Journal of Consciousness Studies* 2:272-88. Reprinted in (J. Shear, ed) *_Explaining Consciousness: The Hard Problem_* (MIT Press, 1999).

Examines a position on which experience is fundamental to the world, and suggests that this ought to lead to panpsychism. With some connections to information and quantum mechanics.

Sellars, R.W. 1960. Panpsychism or evolutionary materialism. *Philosophy of Science* 27:329-49.

Shepherd, J.J. 1974. Panpsychism and parsimony. *Process Studies* 4:3-10.

Sprigge, T.L.S. 1983. The vindication of panpsychism. In *_The Vindication of Absolute Idealism_*. Edinburgh University Press.

van Cleve, J. 1990. Mind -- dust or magic? Panpsychism versus emergence. *Philosophical Perspectives* 4:215-226.

On Nagel 1979: emergence is more plausible than panpsychism. A construal of emergence as nomological supervenience without logical supervenience.

Wright, S. 1953. Gene and organism. *American Naturalist*.

Wright, S. 1977. Panpsychism and science. In (J. Cobb & D. Griffin, eds) *_Mind in Nature_*. University Press of America.

1.4h Intrinsic Monism (Russell, etc) [see also 1.4h]

Blackburn, S. 1992. Filling in space. *Analysis* 52:62-3.

Physics is dispositional, but if there are only bare dispositions, then the world has no nature of its own. And if there are categorical grounds, we have no idea what they could be, except maybe subjective qualia.

Chalmers, D.J. 1996. The metaphysics of information. In *The Conscious Mind*, pp. 301-8 (see also pp. 153-55). Oxford University Press.

An "it from bit" view fits the Russellian metaphysics (described earlier): physics is info from the outside, (proto)experience is info from the inside. The problem is constituting macrophenomenal from microphenomenal; some ideas.

Demopolous, W. & Friedman, M. 1989. The concept of structure in Russell's *The Analysis of Matter*. In (C. Savage & C. Anderson, eds) *Rereading Russell: Essays in Bertrand Russell's Metaphysics and Epistemology*. University of Minnesota Press.

A nice account of Russell's (and Schlick's and Carnap's) structuralism and Newman's objection, with analysis. (N.B. no philosophy of mind.)

Feigl, H. 1958. The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science* 2:370-497. Reprinted (with a postscript) as *The 'Mental' and the 'Physical'*. University of Minnesota Press, 1967.

A long and very interesting essay on the mind-body problem. Ultimately advocates a "structural" view of the physical and identifies experience with the underlying reality, at least for some neurophysiological states.

Feigl, H. 1960. The mind-body problem: Not a pseudo-problem. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

Feigl, H. 1971. Some crucial issues of mind-body monism. *Synthese* 22:295-312.

Feigl, H. 1975. Russell and Schlick: A remarkable agreement on a monistic solution of the mind-body problem. *Erkenntnis* 9:11-34.

Argues that Russell's and Schlick's views on the structural nature of physics and the possible identification of the "content" with experience are quite close to each other. With interesting historical remarks.

Feser, E. 1998. Can phenomenal qualities exist unperceived? *Journal of Consciousness Studies* 4:405-14.

Foster, J. 1982. *The Case for Idealism*. Routledge.

Foster, J. 1991. Lockwood's hypothesis. In *The Immaterial Self*. Oxford University Press.

Argues that the Russellian view is implausible, as the structure of the underlying physical processes does not correspond to the structure and quality of consciousness.

Lockwood, M. 1981. What was Russell's neutral monism? *Midwest Studies in Philosophy* 6:143-58.

Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford University Press. On the mind-body problem and physical theory. Against reductive physicalism; instead, experience is the intrinsic nature of the physical. Explores some potential links with quantum mechanics.

Lockwood, M. 1993. The grain problem. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

On the "grain problem" for the intrinsic-nature view: how do lots of microphysical qualities add up into a smooth experience? Appeals to quantum mechanics and a preferred set of observables.

Lockwood, M. 1998. Unsensed phenomenal qualities: A defence. *Journal of Consciousness Studies* 4:415-18.

Maxwell, G. 1971. Structural realism and the meaning of theoretical terms. *Minnesota Studies in the Philosophy of Science*.

Maxwell, G. 1978. Unity of consciousness and mind-brain identity. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Maxwell, G. 1979. Rigid designators and mind-brain identity. *Minnesota Studies in the Philosophy of Science* 9.

"Brain state" reference is fixed by topic-neutral description; picks out pain but mightn't have, explaining the illusion of contingency. "Nonphysicalist materialism" results, with mind the essence of matter. An important paper.

Newman, M.H.A. 1928. Mr. Russell's causal theory of perception. *Mind*.
Argues against Russell's structuralism: any collection can be arranged to have a given structure, under some relation, so if physics tells us only about structure, it tells us at most the cardinality of the world.

Robinson, H. 1982. Matter: Turning the tables. In *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Rosenberg, G.H. 1997. *A Place for Consciousness: Probing the Deep Structure of the Natural World*. Dissertation, Indiana University.

Russell, B. 1927. *The Analysis of Matter*. London: Kegan Paul.
Argues that physics characterizes the external world only structurally, and leaves intrinsic qualities unspecified. Only experience acquaints us with anything intrinsic. Perhaps the intrinsic nature of physics is experiential?

Schlick, M. 1925. *General Theory of Knowledge*.

Stoljar, D. 1997. Neutral monism. Manuscript.

Stoljar, D. 2001. Two conceptions of the physical. *Philosophy and Phenomenological Research* 62:253-81.

Stubenber, L. 1996. The place of qualia in the world of science. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Stubenber, L. 1997. Austria vs. Australia: Two versions of the identity theory. In (K. Lehrer & J. Marek, eds) *Austrian Philosophy, Past and Present*. Kluwer.

Stubenber, L. 1998. *Consciousness and Qualia*. John Benjamins.

1.5 Consciousness and Content

1.5a Consciousness and Intentionality (Searle, etc)

Cole, D. 1994. Thought and qualia. *Minds and Machines* 4:283-302.

Crane, T. 1998. Intentionality as the mark of the mental. *Philosophy*.

Davies, M. 1995. Consciousness and the varieties of aboutness. In (C. Macdonald, ed) *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford University Press.

Dunlop, C. 2000. Searle's unconscious mind. *Philosophical Psychology* 13:123-148.

Fodor, J. & Lepore, E. 1994. What is the Connection Principle? *Philosophy and Phenomenological Research* 54:837-45.

Searle's formulation of the connection principle is unclear, and there is no formulation is both plausible and interesting.

Georgalis, N. 1996. Awareness, understanding, and functionalism. *Erkenntnis* 44:225-56.

Gillett, E. 1996. Searle and the "deep unconscious". *Philosophy, Psychiatry, and Psychology* 3:191-200.

Gunderson, K. 1990. Consciousness and intentionality: Robots with and without the right stuff. In (C.A. Anderson & J. Owens, eds) *Propositional Attitudes: The Role of Content in Language, Logic, and Mind*. CSLI.

Marbach, E. 1993. *Mental Representation and Consciousness: Toward a Phenomenological Theory of Representation and Reference*. Kluwer.

Meijers, A. 2000. Mental causation and Searle's impossible conception of unconscious intentionality. *International Journal of Philosophical Studies* 8:155-170.

McLoughlin, J. 1999. Unwittingly recapitulating Freud: Searle's concept of a vocabulary of the unconscious. *Ratio* 12:34-53.

Natsoulas, T. 1992. Intentionality, consciousness, and subjectivity. *Journal of Mind and Behavior* 13:281-308.

Nelkin, N. 1989. Propositional attitudes and consciousness. *Philosophy and Phenomenological Research* 49:413-30.

About conscious beliefs. We are not "conscious of" beliefs, merely "conscious that" -- i.e. belief is not phenomenological.

Nelkin, N. 1993. The connection between intentionality and consciousness. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Against Searle: some intentional states aren't even potentially conscious (blindsight, etc) and intentional content doesn't require a particular phenomenal feel. So there's no essential link. With remarks on McGinn.

Schweizer, P. 1994. Intentionality, qualia, and mind/brain identity. *Minds and Machines* 4:259-82.

Seager, W.E. 1999. Conscious intentionality and the anti-Cartesian catastrophe. In *Theories of Consciousness: An Introduction and Assessment*. Routledge.

Searle, J.R. 1984. Intentionality and its place in nature. *Synthese*.
(Subjective) intentionality sure is real. It causes and is caused.

Searle, J.R. 1990. Consciousness, explanatory inversion and cognitive science. *Behavioral and Brain Sciences* 13:585-642.

Advocates a "connection principle": intentional states must be potentially conscious. If not, they're brutally neurophysiological. So cog-sci talk of "intentional" cognitive mechanisms below the conscious level isn't justified.

Searle J.R. 1994. The connection principle and the ontology of the unconscious: A reply to Fodor and Lepore. *Philosophy and Phenomenological Research* 54:847-55.

Clarifying the connection principle -- it's necessary in order to see how certain nonconscious neural states qualify as unconscious mental states.

van Gulick, R. 1988. Consciousness, intrinsic intentionality, and self-understanding machines. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

van Gulick, R. 1995. Why the connection argument doesn't work. *Philosophy and Phenomenological Research* 55:201-7.

van Gulick, R. 1995. How should we understand the relation between intentionality and phenomenal consciousness. *Philosophical Perspectives* 9:271-89.

Worley, S. 1997. Belief and consciousness. *Philosophical Psychology* 10:41-55.
Argues that belief requires consciousness, as we can't make sense of the personal/subpersonal content distinction without appealing to consciousness.

1.5b The Content of Experience

Baldwin, T. 1992. The projective theory of sensory content. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Bermudez, J.L. 1994. Peacocke's argument against the autonomy of nonconceptual representational content. *Mind and Language* 9:402-18.

Bermudez, J. 2000. Naturalized sense data. *Philosophy and Phenomenological Research* 61:353-374.

Bermudez, J.L. 1995. Nonconceptual content: From perceptual experience to subpersonal computational states. *Mind and Language* 10:333-69.

Berger, G. 1987. On the structure of visual sentience. *Synthese* 71:355-70.

Bilgrami, A. 1994. On McDowell on the content of perceptual experience. *Philosophical Quarterly* 44:206-13.

Brewer, B. 1999. *Perception and Reason*. Oxford University Press.

Butchvarov, P. 1980. Adverbial theories of consciousness. *Midwest Studies in Philosophy* 5:261-80.

Church, J. 2000. 'Seeing as' and the double bind of consciousness. *Journal of Consciousness Studies* 7:99-112.

Clark, R. 1973. Sensuous judgments. *Nous* 7:45-56.

Clark, R. 1976. The sensuous content of perception. In (H. Castaneda, ed) *Action, Knowledge, and Reality*. Bobbs-Merrill.

Clark, R. 1981. Sensing, perceiving, thinking. *Grazer Philosophische Studien* 12:273-95.

Crane, T. (ed) 1992. *The Contents of Experience: Essays on Perception*. Cambridge University Press.

Crane, T. 1992. The nonconceptual content of experience. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Davies, W.M. 1996. *Experience and Content: Consequences of a Continuum Theory*. Avebury.

DeBellis, M. 1991. The representational content of musical experience. *Philosophy and Phenomenological Research* 51:303-24.
Contra Peacocke, we don't need sensational properties to understand the content of musical experience. Fine-grained representational properties can do the job, with the help of some Schenkerian analysis.

Gunther, Y.H. 1995. Perceptual content and the subpersonal. *Conference*

6:31-45.

Gunther, Y.H. 2001. Content, illusion, partition. *Philosophical Studies* 102:185-202.

Hamlyn, D.W. 1994. Perception, sensation, and non-conceptual content. *Philosophical Quarterly* 44:139-53.

Jackson, F. 1976. The existence of mental objects. *American Philosophical Quarterly* 13:33-40.

Kelly, S.D. 2001. The non-conceptual content of perceptual experience: Situation dependence and fineness of grain. *Philosophy and Phenomenological Research* 62:601-608.

Kraut, R. 1982. Sensory states and sensory objects. *Nous* 16:277-93.

Langsam, H. 2000. Experiences, thoughts, and qualia. *Philosophical Studies* 99:269-295.

Lowe, E.J. 1992. Experience and its objects. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Martin, M.G.F. 1992. Perception, concepts, and memory *Philosophical Review* 101:745-63.

McDowell, J. 1994. The content of perceptual experience. *Philosophical Quarterly* 44:190-205.

McFarland, D. 1998. Crane on concepts and experiential content. *Analysis* 58:54-58.

Millar, A. 1991. Concepts, experience, and inference. *Mind* 100:495-505.

Natsoulas, T. 1983. What are the objects of perceptual consciousness? *American Journal of Psychology* 96:435-67.

Natsoulas, T. 1994. On the distinction between the object and the content of consciousness. *Journal of Mind and Behavior* 15:239-64.

Noe, A. 1999. Thought and experience. *American Philosophical Quarterly* 36:257-65.

Peacocke, C. 1983. *Sense and Content: Experience, Thought, and their Relations*. Oxford University Press.

Peacocke, C. 1984. Colour concepts and colour experience. *Synthese* 58:365-82.

Peacocke, C. 1989. Perceptual content. In (J.Almog, J. Perry, & H. Wettstein, eds) *Themes from Kaplan*. Oxford University Press.

Peacocke, C. 1992. Scenarios, concepts, and perception. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Peacocke, C. 1994. Nonconceptual content: Kinds, rationales, and relations. *Mind and Language* 4:419-29.

Peacocke, C. 1997. Nonconceptual content defended. *Philosophy and Phenomenological Research*.

Peacocke, C. 2001. Does perception have a nonconceptual content? *Journal of Philosophy* 98:239-264.

Peacocke, C. 2001. Phenomenology and nonconceptual content. *Philosophy and*

Phenomenological Research 62:609-615.

Pendlebury, M. 1987. Perceptual representation. Proceedings of the Aristotelian Society 87:91-106.

Pendlebury, M. 1990. Sense experiences and their contents: A defense of the propositional account. Inquiry 33:215-30.

Lots of reasons why experiences have propositional content (i.e., their content is truth-evaluable, etc). A nice paper.

Schantz, R. 2001. The given regained: Reflections on the sensuous content of experience. Philosophy and Phenomenological Research 62:167-180.

Sedivy, S. 1996. Must conceptually informed perceptual experience involve nonconceptual content? Canadian Journal of Philosophy 26:413-31.

Snowdon, P. 1990. The objects of perceptual experience. Aristotelian Society Supplement, 64:121-50.

Sturgeon, S. 1998. Visual experience. Proceedings of the Aristotelian Society 72:179-200.

Tolhurst, W. 1998. Seemings. American Philosophical Quarterly 35:293-302.

Valberg, J.J. 1992. The Puzzle of Experience. Oxford University Press.

Yoon, B. 2000. Intentionality of perceptual experience. Erkenntnis 52:339-355.

1.5c Representationalism

Beckermann, A. 1995. Visual information-processing and phenomenal consciousness. In (T. Metzinger, ed) Conscious Experience. Ferdinand Schoningh.

Block, N. 1990. Inverted earth. Philosophical Perspectives 4:53-79.
Uses Inverted Earth case, colors and lenses inverted, to argue vs Harman that qualitative states aren't intentional states. Also, less convincingly, to argue that qualitative states aren't functional states.

Block, N. 1996. Mental paint and mental latex. In (E. Villanueva, ed) Perception. Ridgeview.

Block, N. 1999. Sexism, racism, ageism, and the nature of consciousness. Philosophical Topics 26.

Block, N. 2000. Mental paint. In (M. Hahn & M. Ramberg), Essays on Burge. MIT Press.

Dretske, F. 1995. Naturalizing the Mind. MIT Press.

Harman, G. 1990. The intrinsic quality of experience. Philosophical Perspectives.

There are no real qualia problems, just Intentional confusions.

Harman, G. 1996. Explaining objective color in terms of subjective reactions. In (E. Villaneuva, ed) Perception. Ridgeview.

Lalor, B. 1999. Intentionality and qualia. Synthese 121:249-290.

Levine, J. 1997. Are qualia just representations? (Critical notice of Tye) Mind and Language 12:101-13.

Lloyd, D. 1991. Leaping to conclusions: connectionism, consciousness, and the computational mind. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Suggests that conscious states are identical to representational states, and that unconscious representation is impossible; transition between conscious states is non-representational. Appeals to connectionist models in support.

Lloyd, D. 1997. Consciousness and its discontents. *Communication and Cognition* 30:273-284.

Argues that consciousness and representation are distinct, as e.g. the latter depends on context but the former does not.

Lycan, W.G. 1996. Layered perceptual representation. In (E. Villaneuva, ed) *Perception*. Ridgeview.

Lycan, W.G. 1996. *Consciousness and Experience*. MIT Press.

Lycan, W.G. 1998. In defense of the representational theory of qualia. *Philosophical Perspectives* 12:479-87.

McCulloch, G. 1993. The very idea of the phenomenological. *Proceedings of the Aristotelian Society* 67:39-57.

The phenomenological can be reduced to the intentional. Intentional states have a what-it-is-like, and there is no special phenomenal object of introspection.

Neander, K. 1998. The division of phenomenal labor: A problem for representationalist theories of consciousness. *Philosophical Perspectives* 12:411-34.

O'Brien, G. & Opie, J. 1997. Cognitive science and phenomenal consciousness: A dilemma, and how to avoid it. *Philosophical Psychology* 10:269-86.

Rey, G. 1998. A narrow representationalist account of qualitative experience. *Philosophical Perspectives* 12:435-58.

Robinson, W.S. 1998. Intrinsic qualities of experience: Surviving Harman's critique. *Erkenntnis* 47:285-309.

Seager, W.E. 1999. Representational theories of consciousness, parts I and II. In *Theories of Consciousness*. Routledge.

Shoemaker, S. 1990. Qualities and qualia: What's in the mind? *Philosophy and Phenomenological Research Supplement* 50:109-131.

Qualia can't be reduced to standard intentional properties (due to certain inversion cases). Projectivist and sense-reference accounts don't work either. Perhaps qualia are necessarily-illusory intentional properties.

Shoemaker, S. 1991. Qualia and consciousness. *Mind* 100:507-24.

On the relationship between phenomenal and intentional aspects of qualia, and in particular on the accessibility of qualia to conscious awareness. Phenomenal & intentional similarity are connected but must be distinguished.

Stalnaker, R. 1996. On a defense of the hegemony of representation. In (E. Villanueva, ed) *Perception*. Ridgeview.

Sullivan, P.R. 1995. Contentless consciousness and information-processing theories of mind. *Philosophy, Psychiatry, and Psychology* 2:51-59.

Tye, M. 1992. Visual qualia and visual content. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Tye, M. 1994. Do pains have representational content? In (R. Casati, B.

Smith, & S. White, eds) Philosophy and the Cognitive Sciences.
Holder-Pichler-Tempsky.

Tye, M. 1995. A representational theory of pains and their phenomenal character. *Philosophical Perspectives* 9:223-39.
Argues that pain is representational, and that its phenomenal character is narrow nonconceptual content. They have a complex representational structure, with map-like arrays of sentential contents.

Tye, M. 1995. What "what it is like" is like. *Analysis*.
Argues that "what it is like to be X" is an intentional context, which solves some of the associated problems.

Tye, M. 1996. Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind. MIT Press.

Tye, M. 1996. Orgasms again. In (E. Villanueva, ed) Perception. Ridgeview.

Tye, M. 1997. The problem of simple minds: Is there anything it's like to be a honeybee? *Philosophical Studies* 88:289-317.

Tye, M. 1998. Inverted earth, swampman, and representationalism. *Philosophical Perspectives* 12:459-78.

Stoljar, D. 1996. What what it's like isn't like. *Analysis* 56:281-83.

Vinueza, A. 2000. Sensations and the language of thought. *Philosophical Psychology* 13:373-392.

Wager, A. 1999. The extra qualia problem: Synaesthesia and representationism. *Philosophical Psychology* 12:263-281.

Warfield, T. 1999. Against representational theories of consciousness. *Journal of Consciousness Studies* 6:66-69.

White, S. 1994. Color and notional content. *Philosophical Topics* 22:471-503.

1.5d Internalism and Externalism about Experience [see also 2.2]

Davies, M. 1992. Perceptual content and local supervenience. *Proceedings of the Aristotelian Society* 66:21-45.

Argues that perceptual content does not supervene on internal state, even though it is non-conceptual. Constructs an Twin scenario to that effect.
With remarks on the relation between perceptual content and phenomenology.

Davies, M. 1993. Aims and claims of externalist arguments. In (E. Villanueva, ed) Naturalism and Normativity. Ridgeview.

Distinguishes modal and constitutive externalism, characterizes perceptual content and its relation to sensational content, and argues for externalism about perceptual content by examples.

de Vries, W.A. 1996. Experience and the swamp creature. *Philosophical Studies* 82:55-80.

Argues that a swampthing isn't intelligent or intentional, with different physiological processes and no sensations, as these are functional kinds.

Dretske, F. 1996. Phenomenal externalism, or if meanings ain't in the head, where are qualia? In (E. Villanueva, ed) Perception. Ridgeview.

We only have access to qualia through our concepts, which are external; so internal qualia would be inaccessible. So if qualia are knowable, they're external; and if not, why posit them? With comments by Kim, Horwich, Biro.

Forbes, G. 1997. Externalism and scientific Cartesianism. *Mind and Language* 12:196-205.

Kirk, R. 1994. The trouble with ultra-externalism. *Proceedings of the Aristotelian Society* 68:293-307.

Kirk, R. 1996. Why ultra-externalism goes too far. *Analysis* 56:73-79.

Kirk, R. 1998. Consciousness, information, and external relations. *Communication and Cognition* 30:249-71.

McCulloch, G. 1990. Externalism and experience. *Analysis* 50:244-50.
Argues against McGinn that one should embrace a form of "strong externalism" about experience. Experience can be laden with externally-grounded concepts.

McCulloch, G. 1994. Not much trouble for ultra-externalism. *Analysis* 54:265-9.

Sartwell, C. 1995. Radical externalism concerning experience. *Philosophical Studies* 78:55-70.

There is no epistemically available aspect of experience that is determined internally; experiences are "fused" with the environment.

Tappenden, P. 1996. The roundsquare copula: A semantic internalist's rejoinder. *Proceedings of the Aristotelian Society* 96:395-400.

1.5e Miscellaneous

Cam, P. 1984. Consciousness and content-formation. *Inquiry* 27:381-98.

Carruthers, P. 1998. Conscious thinking: Language or elimination?
Mind and Language 13:457-476.

Falk, B. 1993. Consciousness, cognition, and the phenomenal. *Proceedings of the Aristotelian Society* 67:55-73.

On conceptual influences on experience, and aspectual seeing, focusing on bodily and dynamic elements. Self-awareness is not of phenomenal states but *in* them. With commentary by S. Mulhall.

Jacquette, D. 1984. Sensation and intentionality. *Philosophical Studies* 47:229-40.

Sensations don't have intentional objects, they *are* intentional objects.

Maloney, J.C. 1986. Sensuous content. *Philosophical Papers* 15:131-54.

McGinn, C. 1988. Consciousness and content. *Proceedings of the British Academy* 74:219-39. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Comparing the problems of consciousness and content, and reconciling optimism on content with pessimism on consciousness. The phenomenological nature of content may be mysterious, but the individuation of contents is not.

Nelkin, N. 1994. Phenomena and representation. *Philosophy of Science* 45:527-47.

Arguing against the view that phenomenal properties are "read off" in making perceptual judgments. Experiences do not literally have color or shape.

Sosa, E. 1986. Experience and intentionality. *Philosophical Topics* 14:67-83.

On a propositional conception of experience, and making sense of awareness of experience and various problems for sense-data monadicism.

1.6 Aspects of Consciousness

1.6a Self-Consciousness [see also 6.20]

- Anscombe, G.E.M. 1975. The first person. In (S. Guttenplan, ed) *Mind and Language*. Oxford University Press.
- Balaban, O. 1990. *Subject and Consciousness: A Philosophical Inquiry into Self-Consciousness*. Rowman & Littlefield.
- Bealer, G. 1997. Self-consciousness. *Philosophical Review* 106:69-117.
- Bermudez, J. 1997. Reduction and the self. *Journal of Consciousness Studies* 4:458-66.
- Bermudez, J. 1998. *The Paradox of Self-Consciousness*. MIT Press.
- Canfield, J.V. 1990. *The Looking-Glass Self: An Examination of Self-Awareness*. Praeger.
- Campbell, J. 1994. *Past, Space, and Self*. MIT Press.
- Campbell, J. 1995. The body image and self-consciousness. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Cassam, Q. 1995. Transcendental self-consciousness, In (P. Kumar, ed) *The Philosophy of P.F. Strawson*. Indian Council for Philosophical Research.
- Cassam, Q. 1997. *Self and World*. Oxford University Press.
- Castaneda, H. 1989. The reflexivity of self-consciousness: Sameness/identity, data for artificial intelligence. *Philosophical Topics* 17:27-58.
- Cheeks, J.M. & Briggs, S.R. 1982. Self-consciousness and aspects of personality. *Journal of Research in Personality* 16:401-8.
- Chisholm, R.M. 1969. On the observability of the self. *Philosophy and Phenomenological Research* 30:7-21.
- Christofidou, A. 2000. Self-consciousness and the double immunity. *Philosophy* 75:539-570.
- Church, J. 1990. Judgment, self-consciousness, and object-independence. *American Philosophical Quarterly* 27:51-60.
- Davis, L.H. 1989. Self-consciousness in chimps and pigeons. *Philosophical Psychology* 2:249-59.
- Delius, H. 1981. *Self-Awareness: A Semantical Inquiry*. Beck.
- Dennett, D.C. 1992. The self as the center of narrative gravity. In (F. Kessel, P. Cole, & D.L. Johnson, eds) *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.
- Eilan, N. 1995. Consciousness and the self. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Eilan, N. Marcel, A.J. & Bermudez, J. 1995. Self-consciousness and the body: Interdisciplinary issues. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Falk, A. 1995. Consciousness and self-reference. *Erkenntnis* 43:151-80.

Frith, U. & Happe, F. 1999. Theory of mind and self-consciousness: What is it like to be autistic? *Mind and Language* 14:1-22.

Gallagher, S. & Meltzoff, A. 1996. The earliest sense of self and others: Merleau-Ponty and recent developmental studies. *Philosophical Psychology* 9:211-33.

Gallagher, S. 1996. The moral significance of primitive self-consciousness: A response to Bermudez. *Ethics* 107:129-40.

Gallagher, S. 2000. Self-reference and schizophrenia: A cognitive model of immunity to error through misidentification. In (D. Zahavi, ed) *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience*. John Benjamins.

Ganerri, J. 2000. Cross-modality and the self. *Philosophy and Phenomenological Research* 61:639-657.

Gennaro, R.J. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333-47.
Argues that consciousness entails having episodic memory, which entails self-consciousness. So consciousness entails self-consciousness.

Gennaro, R.J. 1996. *Consciousness and Self-consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*. John Benjamins.

Hurley, S.L. 1998. Nonconceptual self-consciousness and agency: Perspective and access. *Communication and Cognition* 30:207-247.

James, W. 1890. The consciousness of self. In *The Principles of Psychology*.

Laycock, S.W. 1998. Consciousness it/self. *Journal of Consciousness Studies* 5:141-152.

Martin, M.G.F. 1995. Bodily awareness: A sense of ownership. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.

McCullagh, M. 2000. Functionalism and self-consciousness. *Mind and Language* 15:481-499.

Meijsing, M. 2000. Self-consciousness and the body. *Journal Of Consciousness Studies* 7:34-50.

Metzinger, T. 2000. The subjectivity of subjective experience: A representationist analysis of the first-person perspective. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Mittal, K.K. 1979. Self-identity and self-consciousness. *Indian Philosophical Quarterly* 7:159-63.

Neisser, U. 1991. Five kinds of self-knowledge. *Philosophical Psychology* 1:35-59.

O'Hear, A. 1989. Evolution, knowledge, and self-consciousness. *Inquiry* 32:127-150.

Parker, S.T., Mitchell, R.M., & Boccia, M.L. 1994. *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Proust, J. 2000. Awareness of agency: Three levels of analysis. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Richards, W. 1984. Self-consciousness and agency. *Synthese* 61:149-71.
Self-consciousness is consciousness of agency. Castaneda/Nozick/Nagel.

Shoemaker, S. 1968. Self-reference and self-awareness. *Journal of Philosophy* 65:555-67.

Shoemaker, S. 1986. Introspection and the self. *Midwest Studies in Philosophy*.

Shoemaker, S. 1994. Self-knowledge and "inner sense". *Philosophy and Phenomenological Research* 54:249-314.

Smith, D.W. 1986. The structure of (self-) consciousness. *Topoi* 5:149-56.

Sosa, E. 1983. Consciousness of self and of the present. In (J. Tomberlin, ed) *_Agent, Language, and the Structure of the World_*. Hackett.

Stephens, G. L. & Graham, G. 1994. Self-consciousness, mental agency, and the clinical psychopathology of thought-insertion. *Philosophy, Psychiatry, and Psychology* 1:1-10.

Stephens, G.L. & Graham, G. 2000. *_When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts_*. MIT Press.

Strawson, P.F. 1974. Self, mind, and body. In *_Freedom and Resentment and Other Essays_*.

Strawson, G. 1997. 'The self'. *Journal of Consciousness Studies* 4:405-28.

van Gulick, R. 1988. A functionalist plea for self-consciousness. *Philosophical Review* 97:149-88.

How functionalism can handle consciousness: Self-consciousness is the possession of reflexive metapsychological information. This helps understand learning, representation, and belief. Phenomenal experience is still tricky.

Varela, F.G. 1971. Self-consciousness: Adaptation or epiphenomenon? *Stud Gen* 24:426-39.

White, S. 1987. What is it like to be a homunculus? *Pac Philosophical Quarterly* 68:148-74.

Weird examples of homunculi that are conscious but not self-conscious.
Self-consciousness, not consciousness, is what really counts.

Zahavi, D. 2000. Self and consciousness. In (D. Zahavi, ed) *_Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience_*. John Benjamins.

1.6b The Unity of Consciousness [see also 5.12d, 6.1e, 6.1g]

Arvidson, P. 2000. Transformations in consciousness: Continuity, the self and marginal consciousness. *Journal Of Consciousness Studies* 7:3-26.

Bayne, T. 2000. The unity of consciousness: Clarification and defence. *Australasian Journal Of Philosophy* 78:248-254.

Beahrs, J.O. 1982. *_Unity and Multiplicity: Multilevel Consciousness of Self in Hypnosis, Psychiatric Disorder, and Mental Health_*. Brunner/Mazel.

Beahrs, J.O. 1983. Co-consciousness: A common denominator in hypnosis, multiple personality, and normalcy. *American Journal of Clinical Hypnosis* 26:100-13.

Brooks, D.H.M. 1985. Strawson, Hume, and the unity of consciousness. *Mind* 94:583-86.

Brooks, D.H.M. 1995. *_The Unity of the Mind_*. St. Martin's Press.

- Cotterill, R.M.J. 1995. On the unity of conscious experience. *Journal of Consciousness Studies* 2:290-311.
- Dainton, B. 2000. *_Stream of Consciousness: Unity and continuity in conscious experience_*. Routledge.
- Eccles, J. 1985. *_The Brain and the Unity of Conscious Experience_*. Cambridge University Press.
- Feinberg, T. 2000. The nested hierarchy of consciousness: A neurobiological solution to the problem of mental unity. *Neurocase* 6:75-81.
- Fox, I. 1985. The individualization of consciousness. *Philosophical Topics* 13:119-43.
- Hamlyn, D.W. 1996. The unity of the senses and self-consciousness. In *_Understanding Perception: The Concept and its Conditions_*. Avebury Press.
- Hill, C.S. 1991. Unity of consciousness, other minds, and phenomenal space. In *_Sensations: A Defense of Type Materialism_*. Cambridge University Press.
- Humphrey, N. 2000. One-self: A meditation on the unity of consciousness. *Social Research* 67:1059-1066.
- Hurley, S. 1994. Unity and objectivity. In (C. Peacocke, ed) *_Objectivity, Simulation, and the Unity of Consciousness_*. Oxford University Press.
- Hurley, S. 1998. Unity, neuropsychology, and action. In *_Consciousness in Action_*. Harvard University Press.
- James, W. 1895. The knowing of things together. *Psychological Review* 2:105-24.
- Kobes, B. 2000. Unity of consciousness and bi-level externalism. *Mind and Language* 15:528-544.
- Lockwood, M. 1994. Issues of unity and objectivity. In (C. Peacocke, ed) *_Objectivity, Simulation, and the Unity of Consciousness_*. Oxford University Press.
- Malpas, J. 1999. Constituting the mind: Kant, Davidson, and the unity of consciousness. *International Journal of Philosophical Studies* 7:1-30.
- Marcel, A.J. 1993. Slippage in the unity of consciousness. In *_Experimental and Theoretical Studies of Consciousness_* (Ciba Foundation Symposium 174). Wiley.
- Marcel, A.J. 1994. What is relevant to the unity of consciousness? In (C. Peacocke, ed) *_Objectivity, Simulation, and the Unity of Consciousness_*. Oxford University Press.
- Marks, C. 1980. *_Commissurotomy, Consciousness, and Unity of Mind_*. MIT Press.
- Marks, L.E. 1978. *_The Unity of the Senses: Interrelations among the Modalities_*. Academic Press.
- Maxwell, G. 1978. Unity of consciousness and mind-brain identity. In (J. Eccles, ed) *_Mind and Brain_*. Paragon House.
- McInerney, P.K. 1985. Person-stages and unity of consciousness. *American Philosophical Quarterly* 22:197-209.

Metzinger, T. 1995. Faster than thought: Holism, homogeneity, and temporal coding. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22:396-413. Reprinted in *_Mortal Questions_* (Cambridge University Press, 1979).

Natsoulas, T. 1979. The unity of consciousness. *Behaviorism* 7:45-63.

Natsoulas, T. 1984. Concerning the unity of consciousness: I. Varieties of conscious unity. *Imagination, Cognition and Personality* 3:281-303.

Natsoulas, T. 1986. Concerning the unity of consciousness: II. William James on personal conscious unity. *Imagination, Cognition abd Personality* 5:21-30.

Oakley, D.A. & Eames, L.C. 1986. The plurality of consciousness. In (D. Oakley, ed) *_Mind and Brain_*. Methuen.

O'Brien, G. & Opie, J. 1998. The disunity of consciousness. *Australasian Journal of Philosophy* 76:378-95.

O'Brien, G. & Opie, J. 2000. Disunity defended: A reply to Bayne. *Australasian Journal Of Philosophy* 78:255-263.

O'Shaughnessy, B. 1994. The diversity and unity of action and perception. In (T. Crane, ed) *_The Contents of Experience_*. Cambridge University Press.

Rosenberg, G.H. 1998. The boundary problem for phenomenal individuals. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness 1996_*. MIT Press.

Rossman, N.I. 1991. *_Consciousness: Separation and Integration_*. SUNY Press.

Schleichert, H. 1985. On the concept of unity of consciousness. *Synthese* 64:411-20.

Shoemaker, S. 1996. Unity of consciousness and consciousness of unity. In *_The First-Person Perspective and Other Essays_*. Cambridge University Press.

Stevenson, L. 2000. Synthetic unities of experience. *Philosophy and Phenomenological Research* 60:281-306.

Tinnin, L. 1990. Mental unity, altered states of consciousness, and dissociation. *Dissociation: Progress in the Dissociative Disorders* 3:154-59.

von der Malsburg, C. 1997. The coherence definition of consciousness. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *_Cognition, Computation, and Consciousness_*. Oxford University Press.

Ward, A. 1980. Materialism and the unity of consciousness. *Analysis* 40:144-46.

Watkins, J.W.N. 1982. A basic difficulty in the mind-brain identity hypothesis. In (J. Eccles, ed) *_Mind and Brain_*. Paragon House.

Zohar, D. 1995. A quantum-mechanical model of consciousness and the emergence of 'I'. *Minds and Machines* 5:597-607.

1.6c Homogeneity of Consciousness (Sellars, etc)

Clark, A. 1989. The particulate instantiation of homogeneous pink. *Synthese* 80:277-304.

Explains homogeneity in terms of nontransitive matching among pixelized parts

of vision. Experience of continuity, not continuous experience. Experiences may have subphenomenal parts (e.g. invisible pixels).

Cornman, J.W. 1970. Sellars, scientific realism, and *sensa*. Review of *Metaphysics* 23:417-51.

Delaney, C.F. 1971. Sellars' grain argument. *Australasian Journal of Philosophy* 50:14-16.

Friedman, I.S. 1989. Ultimate homogeneity: A dialogue. *Philosophy Research Archives* 14:425-53.

Gunderson, K. 1974. The texture of mentality. In (R. Bambrough, ed) *Wisdom: Twelve Essays*. Blackwell.

Lockwood, M. 1993. The grain problem. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

Lycan, W.G. 1987. Sellars' "grain" argument. In *Consciousness*.

Metzinger, T. 1995. Faster than thought: Holism, homogeneity, and temporal coding. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Richardson, R.C. & Muilenberg, G. 1982. Sellars and sense impressions. *Erkenntnis*.

Sellars, W.S. 1963. Philosophy and the scientific image of man. In *Science, Perception, and Reality*. Humanities Press/Ridgeview.

Sellars, W.S. 1971. Seeing, sense impressions, and *sensa*: A reply to Cornman. Review of *Metaphysics*.

1.6d Knowledge of Consciousness [see also 1.3a, 1.8a, 5.13, 6.2i]

Alston, W.P. 1971. Varieties of priveleged access. *American Philosophical Quarterly* 8:223-41.

Alston, W.P. 1983. What's wrong with immediate knowledge? *Synthese* 55:73-96.

Armstrong, D.M. 1963. Is introspective knowledge incorrigible? *Philosophical Review* 62:417-32.

Armstrong, D.M. 1976. Incorrigibility, materialism, and causation. *Philosophical Studies* 30:125-28.

Bayne, T. 2001. Chalmers on the justification of phenomenal judgments. *Philosophy and Phenomenological Research* 62:407-19.

Bradley, R.D. 1964. Avowals of immediate experience. *Mind* 73:186-203.

Chalmers, D.J. 2002. The content and epistemology of phenomenal belief. In (Q. Smith & A. Jokic, eds) *Aspects of Consciousness*. Oxford University Press.

Chandler, J.H. 1970. Incorrigibity and classification. *Australasian Journal of Philosophy* 48:101-6.

Conee, E. 1994. Phenomenal knowledge. *Australasian Journal of Philosophy*.
Mary knew all the facts about qualia beforehand, she just wasn't acquainted with them. Knowledge by acquaintance isn't factual knowledge.

Dretske, F. 1999. The mind's awareness of itself. *Philosophical Studies* 95:103-24.

- Dunlop, C.E.M. 1977. Lehrer and Ellis on incorrigibility. *Australasian Journal of Philosophy* 55:201-5.
- Echelbarger, C.G. 1981. An alleged legend. *Philosophical Studies* 39:227-46.
- Ellis, B. 1976. Avowals are more corrigible than you think. *Australasian Journal of Philosophy* 55:201-5.
- Francescotti, R.M. 1994. Qualitative beliefs, wide content, and wide behavior. *Nous* 28:396-404.
Qualitative beliefs can supervene on behavioral dispositions even if absent/
/inverted qualia are possible. We just individuate belief contents and
behavior widely, with wide content fixed to the qualia.
- Hill, C.S. 1988. Introspective awareness of sensations. *Topoi* 7:11-24.
- Imlay, R.A. 1969. Immediate awareness. *Dialogue* 8:228-42.
- Jackson, F. 1973. Is there a good argument against the incorrigibility thesis? *Australasian Journal of Philosophy* 51:51-62.
- Kirk, R. 1971. Armstrong's analogue of introspection. *Philosophical Quarterly* 21:158-62.
- Kornblith, H. 1998. What is it like to be me? *Australasian Journal of Philosophy* 76:48-60.
- Leeds, S. 1993. Qualia, awareness, Sellars. *Nous* 27:303-330.
A discussion of in what sense we are aware of qualia, and how we can have
beliefs about them, with reference to Sellars. Ends up reducing qualia to
phenomenal beliefs in a language of thought. A rich and subtle paper.
- Levin, J. 2001. The myth of Jones and the return of subjectivity. *Mind and Language* 16:173-192.
- MacDonald, C. 1999. Shoemaker on self-knowledge and inner sense. *Philosophy and Phenomenological Research* 59:711-38.
- Mackie, J.L. 1963. Are there any incorrigible empirical statements? *Australasian Journal of Philosophy* 41:12-28.
- Margolis, J. 1970. Indubitability, self-intimating states, and privileged access. *Journal of Philosophy* 67:918-31.
- Martin, M. 1998. An eye directed outward. In (C. Wright, B. Smith, and C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.
- Nakhtnikian, G. 1968. Incorrigibility. *Philosophical Quarterly* 18:207-15.
- Nida-Rumelin, M. 1995. What Mary couldn't know: Belief about phenomenal states. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.
Reconstructing the Mary case as a Marianna case, and introducing a
distinction between phenomenal and nonphenomenal beliefs, to support the
knowledge argument. A very nice and sophisticated paper.
- Nida-Rumelin, M. 1997. On belief about experiences: An epistemological distinction applied to the knowledge argument. *Philosophy and Phenomenological Research*.
- Odegard, D. 1992. Inner states. *Personalist Forum* 8:265-73.
- Pappas, G. 1974. Incorrigibility, knowledge, and justification. *Philosophical*

Studies 25:219-25.

Pappas, G. 1976. Incorrigibility and central-state materialism. *Philosophical Studies* 29:445-56.

Parsons, K.P. 1970. Mistaking sensations. *Philosophical Review*.

Peacocke, C. 1998. Conscious attitudes, attention, and self-knowledge. In (C. Wright, B. Smith, and C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Pollock, J.L. 1970. Perceptual knowledge. *Philosophical Review* 80:287-319.

Robinson, W.S. 1975. The legend of the given. In (H. Castaneda, ed) *Action, Knowledge, and Reality*. Bobbs-Merrill.

Robinson, W.S. 1982. Causation, sensation, and knowledge. *Mind* 91:524-40.
Argues that we can have non-inferential knowledge of sensations even if they don't make a causal difference to our beliefs, so epiphenomenalism is OK.
All theories have a similar problems.

Rosenberg, J. 2000. Perception vs. inner sense: A problem about direct awareness. *Philosophical Studies* 101:143-160.

Schick, T.W. 1992. The epistemic role of qualitative content. *Philosophy and Phenomenological Research* 52:383-93.
Contra Sellars, Rorty, and Churchland: knowledge of qualitative content is an important aspect of our understanding of mental concepts, although it is not everything.

Sellars, W. 1956. Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science* 1:253-329. Reprinted as *Empiricism and the Philosophy of Mind*. Harvard University Press, 1997.

Shoemaker, S. 1990. First-person access. *Philosophical Perspectives* 4:187-214.
We have a limited special authority about the contents of our mental states. This follows from the link between a state and beliefs about it in the functional definition of that kind of state.

Smart, J.J.C. 1962. Brain processes and incorrigibility. *Australasian Journal of Philosophy* 40:68-70.
Reply to Baier 1962: epistemological authority is compatible with materialism. Mental state reports are not completely incorrigible, though.

Solomon, R.C. 1975. Minimal incorrigibility. *Australasian Journal of Philosophy* 53:254-56.

Sprigge, T.L.S. 1981. Knowledge of subjectivity. *Theoria to Theory* 14:313-25.

Tibbetts, P. 1972. Feigl on raw feels, the brain, and knowledge claims: Some problems regarding theoretical concepts. *Dialectica* 26:247-66.

Tomberlin, J.E. 1975. A problem with incorrigibility. *Philosophia* 5:507-12.

von Eckardt, B. 1975. Some consequences of knowing everything (essential) there is to know about one's mental states. *Review of Metaphysics* 29:3-18.

Wallraff, C.F. 1953. On immediacy and the contemporary dogma of sense-certainty. *Journal of Philosophy*.

Warner, R. 1993. Incorrigibility. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

Warner, R. 1994. In defense of a dualism. In (R. Warner & T. Szubka, eds) The Mind-Body Problem: A Guide to the Current Debate. Blackwell.

Warner, R. 1996. Facing ourselves: Incorrighibility and the mind-body problem. *Journal of Consciousness Studies* 3:217-30.

1.6e The Function of Consciousness [see also 1.4e, 6.4a]

Baars, B. 1988. The functions of consciousness. In A Cognitive Theory of Consciousness. Cambridge University Press.

Argues for nine major functions of consciousness: in defining inputs, adaptation, debugging, recruiting & control, prioritizing, decision-making, analogy-forming, self-monitoring, and self-maintenance.

Banks, W.P. 1996. How much work can a quale do? *Consciousness and Cognition* 5:368-80.

Bechtel, W. & Richardson, R.C. 1983. Consciousness and complexity: evolutionary perspectives on the mind-body problem. *Australasian Journal of Philosophy* 61:378-95.

Contra Popper, evolution doesn't provide an argument against physicalism or epiphenomenalism. Speculation on what the function of consciousness might be, and how it might be realized: e.g. selecting information.

Block, N. 1995. On a confusion about the function of consciousness. *Behaviora and Brain Sciences* 18:227-47.

Separates phenomenal consciousness from access consciousness, and argues that cases like blindsight only suggest a function for access consciousness, not phenomenal consciousness. The latter remains a mystery. With commentaries.

Bringsjord, S. & Noel, R. 1998. Why did evolution engineer consciousness? In (G. Mulhauser, ed) Evolving Consciousness.

DeLancey, C. 1996. Emotion and the function of consciousness. *Journal of Consciousness Studies* 3:492-99.

Dretske, F. 1997. What good is consciousness? *Canadian Journal of Philosophy* 27:1-15.

Flanagan, O. & Polger, T. 1995. Zombies and the function of consciousness. *Journal of Consciousness Studies* 2:313-21.

Argues for the possibility of zombies (contra Moody), then notes that any function could be performed by an unconscious zombie, it seems, so there's no function of consciousness in sight.

Flanagan, O. & Polger, G. 1998. Consciousness, adaptation, and epiphenomenalism. In (G. Mulhauser, ed) Evolving Consciousness.

Gregory, R.L. 1996. What do qualia do? *Perception* 25:377-79.

Suggests that qualia serve to distinguish hypotheses about present from past.

Kraemer, E.R. 1984. Consciousness and the exclusivity of function. *Mind* 93:271-5.

Contra Mott 1982: Function needn't be exclusive, and brain processes and consciousness may share a function, due to their close relationship.

McGinn, C. 1981. A note on functionalism and function. *Philosophical Topics* 12:169-70.

Function always underdetermines intrinsic nature, so absent/inverted qualia cases aren't incompatible with consciousness having a function.

Mott, P. 1982. On the function of consciousness. *Mind* 91:423-9.

Consciousness doesn't have a function, as any function it might have is a function of brain processes.

Perlis, D. 1997. Consciousness as self-function. *Journal of Consciousness Studies* 4:509-25.

Place, U. T. 2000. The causal potency of qualia: Its nature and its source. *Brain and Mind* 1:183-192.

Popper, K.R. 1978. Natural selection and the emergence of mind. *Dialectica* 32:339-55.

Ramachandran, V.S. & Hirstein, W. 1998. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies* 4:429-57.

Shanon, B. 1998. What is the function of consciousness? *Journal of Consciousness Studies* 5:295-308.

Tye, M. 1996. The function of consciousness. *Nous* 30:287-305.

Argues that the function of consciousness is not obvious, but that once one accepts a representational view of consciousness, it becomes obvious.

van Gulick, R. 1989. What difference does consciousness make? *Philosophical Topics* 17:211-30.

Trying to counter absent qualia arguments by finding a role for consciousness e.g. in metacognition, or as a way to achieve semantic transparency. But consciousness doesn't seem necessary for these, so it's still a mystery.

van Gulick, R. 1994. Deficit studies and the function of phenomenal consciousness. In (G. Graham & G.L. Stephens, eds) *Philosophical Psychopathology*. MIT Press.

Velmans, M. 1992. Is human information-processing conscious? *Behavioral and Brain Sciences* 14:651-69.

Uses experimental evidence to argue that consciousness is functionally inessential: the tasks associated with consciousness can be performed without consciousness. Only focal-attentive processing is required.

1.7 Qualia

1.7a General

Burgess, J.A. 1990. Phenomenal qualities and the nontransitivity of matching. *Australasian Journal of Philosophy*.

Clark, Andy. 2000. A case where access implies qualia? *Analysis* 60:30-37.

Clark, A. 1985. Qualia and the psychophysical explanation of color perception. *Synthese* 65:377-405.

One can give an information-theoretic explanation of color perception, which leaves nothing out. Rebuts various qualia objections, e.g. from the possibility of inversion. Qualia are codes for external properties.

Clark, A. 1992. *Sensory Qualities*. Clarendon.

Argues that psychology is in the business of explaining sensory qualities, and does a perfectly good job using discriminability as a basis. With detailed argument and many interesting examples.

Clark, A. 2000. *A Theory of Sentience*. Oxford University Press

Cunningham, B. 2001. Capturing qualia: Higher-order concepts and connectionism. *Philosophical Psychology* 14:29-41.

Fox, I. 1989. On the nature and cognitive function of phenomenal content -- Part one. *Philosophical Topics* 17:81-103.

Searching for a theory of qualia: rejects epiphenomenalism, separation of the form and quality of experience, and immediate perception of phenomenal objects. Experience consists in represented (inexistent) objects of thought.

Gilbert, P. 1992. Immediate experience. *Proceedings of the Aristotelian Society* 66:233-250.

Against an account of phenomenal content as given by inner discrimination. Argues that the character of experience consists in its reason-giving role.

Gustafson, D. 1998. Pain, qualia, and the explanatory gap. *Philosophical Psychology* 11:371-387.

Hubbard, T.L. 1996. The importance of a consideration of qualia to imagery and cognition. *Consciousness and Cognition* 3:327-58.

Jakab, Z. 2000. Ineffability of qualia: A straightforward naturalistic explanation. *Consciousness and Cognition* 9:329-351.

Kind, A. 2001. Qualia realism. *Philosophical Studies* 104:143-62.

Kitcher, P.S. 1979. Phenomenal qualities. *American Philosophical Quarterly* 16:123-9.

Qualia problems stem from assuming direct awareness of perceptual states. Instead, we should acknowledge only an ability to detect and label these states. Also argues for the possibility of unconscious and illusory pains.

Leeds, S. 1993. Qualia, awareness, Sellars. *Nous* 27:303-330.

A discussion of in what sense we are aware of qualia, and how we can have beliefs about them, with reference to Sellars. Ends up reducing qualia to phenomenal beliefs in a language of thought. A rich and subtle paper.

Leon, M. 1988. Characterising the senses. *Mind and Language* 3:243-70.

Levine, J. 1995. Qualia: Intrinsic, relational, or what? In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Lormand, E. 1995. Qualia! (Now showing at a theater near you.) *Philosophical Topics* 22:127-156.

Mandik, P. 1999. Qualia, space, and control. *Philosophical Psychology* 12:47-60.

Nelkin, N. 1987. How sensations get their names. *Philosophical Studies* 51:325-39.

Sensations are an inessential element of experiences. Experiences are typed by their cognitive component, and the naming of sensations is derivative on this. With examples and empirical evidence about pain, color, perception.

Nelkin, N. 1990. Categorizing the senses. *Mind and Language*.

Putnam, H. 1981. Mind and body. In *_Reason, Truth and History_*. Cambridge University Press.

Considers qualia, inverted and absent, and various other stuff. Wishy-washy.

Rey, G. 1993. Sensational sentences. In (M. Davies & G. Humphreys, eds) *_Consciousness: Philosophical and Psychological Essays_*. Blackwell.

Explicating sensory experience in terms of an appropriate computational relation to a sentence in the language of thought. Argues that this handles

many features of qualia (privacy, ineffability, grainlessness, unity, etc).

Robinson, W.S. 1999. Qualia realism and neural activation patterns. *Journal of Consciousness Studies* 10:65-80.

Shepard, R.N. 1993. On the physical basis, linguistic representation, and conscious experience of colors. In (G. Harman, ed) *Conceptions of the Human Mind: Essays in Honor of George A. Miller*. Lawrence Erlbaum.

Shoemaker, S. 1975. Phenomenal similarity. *Critica* 7:3-37. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).
Where does similarity come from? From belief therein? Similarity of experience = experience of similarity. Also relation to projectibility.

Shoemaker, S. 1994. Phenomenal character. *Nous* 28:21-38.
Phenomenal character is bestowed by representation of certain relational properties, defined by relation to experience. With a discussion of possible candidates, and argument against other views such as projectivism.

Sleutels, J. 1998. Phenomenal consciousness: Epiphenomenalism, naturalism and perceptual plasticity. *Communication and Cognition* 31:21-55.

1.7b Qualia and Materialism [see also 1.2, 1.3]

Clark, A. 1985. A physicalist theory of qualia. *Monist* 68:491-506.
A Goodman-like theory of qualia discrimination.

Cornman, J.W. 1971. *Materialism and Sensations*. Yale University Press.

Double, R. 1985. Phenomenal properties. *Philosophy and Phenomenological Research* 45:383-92.
A somewhat vague defense of materialism against objections from phenomenal properties. The only problems are epistemological.

Harding, G. 1991. Color and the mind-body problem. *Review of Metaphysics* 45:289-307.
On the unique nature of color expanses, which are laid bare to perception as they are in themselves. These are incompatible with functionalist accounts of mind, but might still be physical, on a broader conception thereof.

Holborow, L.C. 1973. Materialism and phenomenal qualities. *Aristotelian Society Supplement* 47:107-19.

Horgan, T. 1987. Supervenient qualia. *Philosophical Review* 96:491-520.
Arguing from the causal efficacy of qualia and the closedness of physical causation to the conclusion that qualia conceptually supervene on the physical. A very thorough paper.

Jolley, K.D. & Watkins, M. 1998. What is it like to be a phenomenologist? *Philosophical Quarterly* 48:204-9.
A reply to Raffman 1995. Maybe our experiences are no more fine-grained than our concepts. Even our experiences of unique hues may be coarse.

Lewis, D. 1995. Should a materialist believe in qualia? *Australasian Journal of Philosophy* 73:140-44.
Materialists can believe in qualia, qua occupier of the folk psychological role. But they cannot accept the Identification Thesis, that having qualia allows us to know exactly what they are.

Lycan, W.G. 1988. Phenomenal objects: A backhanded defense. *Philosophical Perspectives* 3:513-26.
Argues that qualia, if viewed as simple properties of phenomenal individuals,

are problematic for materialism. Considers the case for phenomenal individuals, and argues that they are intentional inexistent.

Marras, A. 1993. Materialism, functionalism, and supervenient qualia. Dialogue 32:475-92.

Qualia aren't reducible to physical properties, but they are supervenient (and ontologically dependent) on microfunctional properties. With remarks on the knowledge argument, Kripke, absent qualia, epiphenomenalism, etc.

Mellor, D.H. 1973. Materialism and phenomenal qualities II. Aristotelian Society Supplement 47:107-19.

Raffman, D. 1995. On the persistence of phenomenology. In (T. Metzinger, ed) Conscious Experience. Ferdinand Schoningh.

Argues that our inability to reidentify qualia is a problem for materialism. How are they represented? Empty demonstrative would be vacuous, predicate would be reidentified, so maybe a plain presentation? Very interesting.

Tye, M. 1986. The subjective qualities of experience. Mind 95:1-17.

Absent/inverted qualia aren't really imaginable. The Knowledge Argument fails, as discovering new experiences doesn't imply learning new facts, but only coming to know old facts in a new way.

1.7c Eliminativism about Qualia [see also 1.4d]

Dennett, D.C. 1978. Why you can't make a computer that feels pain. Synthese 38. Reprinted in Brainstorms* (MIT Press, 1978).

The concept of pain is incoherent, as it's asked to do too many things at once. With a discussion of drugs, flowcharts, reportability, etc.

Dennett, D.C. 1981. Wondering where the yellow went. Monist 64:102-8.

A response to Sellars. All there is to seeing occurrent yellow is the judgment that one is seeing occurrent yellow.

Dennett, D.C. 1988. Quining qualia. In (A. Marcel & E. Bisiach, eds) Consciousness in Contemporary Science. Oxford University Press.

Argues against the existence of ineffable, intrinsic, private, directly accessible properties. With lots of meaty-thought experiments, and arguments that there is no fact of the matter about inversion cases.

Dennett, D.C. 1991. Lovely and suspect qualities. In (E. Villanueva, ed) Consciousness. Ridgeview.

Everett, A. 1996. Qualia and vagueness. Synthese 106:205-226.

There are no qualia: qualia would have to be vague (for Sorites reasons), but there can be no vague properties in nature. The usual Sorites defenses don't work here, as there's no appearance/reality distinction for qualia.

Jacoby, H. 1985. Eliminativism, meaning, and qualitative states. Philosophical Studies 47:257-70.

Arguing against eliminativism for qualia. Even if nothing satisfies all the common-sense properties of qualia, reference of qualia terms is still fixed under a Putnam-style theory of meaning. Argues for scientific functionalism.

Levin, M. 1981. Phenomenal properties. Philosophy and Phenomenological Research 42:42-58.

There are no irreducible phenomenal properties. Materialism can handle our direct awareness of inner states by the right sort of causal connection. Gives a materialism account of discrimination and learning mental concepts.

Levine, J. 1994. Out of the closet: A qualophile confronts qualophobia. Philosophical Topics 22:107-126.

On bold vs. modest qualophilia, and against various qualophobic strategies.
With remarks on scientific objectivity, qualia as an explanandum, and on how
our knowledge of qualia is consistent with the conceivability of zombies.

Ross, D. 1993. Quining qualia Quine's way. Dialogue 32:439-59.

Seager, W.E. 1993. The elimination of experience. Philosophy and
Phenomenological Research 53:345-65.

Dennett's 1988 argument against ineffability, etc., doesn't nearly make the
case against qualia, and largely relies on verificationist assumptions.

Wright, E.W. 1989. Querying "Quining Qualia". Acta Analytica 4:9-32.

1.7d The Inverted Spectrum

Block, N. 1990. Inverted earth. Philosophical Perspectives 4:53-79.

Uses Inverted Earth case, colors and lenses inverted, to argue vs Harman
that qualitative states aren't intentional states. Also, less convincingly,
to argue that qualitative states aren't functional states.

Campbell, N. 2000. Physicalism, qualia inversion, and affective states.
Synthese 124:239-256.

Casati, R. 1990. What is wrong in inverting spectra? Teoria 10:183-6.

Churchland, P.M. & Churchland, P.S. 1981. Functionalism, qualia and
intentionality. Philosophical Topics 12:121-32. Reprinted in _A
Neurocomputational Perspective_ (MIT Press, 1989).

Functional role counts more than qualitative content in determining what
e.g. "redness" is.

Clark, A. 1985. Spectrum inversion and the color solid. Southern Journal of
Philosophy 23:431-43.

Argues that there could be inverted spectra even without a symmetrical color
space. Qualia must be distinguished from their place in color space.

Cole, D.J. 1990. Functionalism and inverted spectra. Synthese 82:207-22.

Acquired spectrum inversions do not refute functionalism, if qualia revert
after behavioral adaptation (as they do with inverting lenses).

Dennett, D.C. 1994. Instead of qualia. In (A. Revonsuo & M. Kamppinen, eds)
Consciousness in Philosophy and Cognitive Neuroscience. Lawrence Erlbaum.

Describes some "inverted spectrum" scenario in computer registers, and argues
that in the absence of a "central clearing house", the inversion of qualia
is indeterminate. There's no reason to believe in non-dispositional qualia.

Gert, B. 1965. Imagination and verifiability. Philosophical Studies 16:44-47.

Inverted spectra with constant behavior is a meaningful hypothesis even under
verificationism. Switching nerve endings, tinting contact lenses, etc.

Hardin, C.L. 1987. Qualia and materialism: Closing the explanatory gap.
Philosophy and Phenomenological Research 48:281-98.

On the physiological bases of phenomenal states, particularly color.
Inverted spectrum isn't really coherent, as coolness/warmth would have to
be inverted too. So the contingency of qualia is diminished.

Hardin, C.L. 1988. _Color for Philosophers_. Hackett.

Distinguishes various functionally distinct inverted spectrum cases.

Hardin, C.L. 1991. Reply to Levine. Philosophical Psychology 4:41-50.

Reply to Levine 1991. "Green residue" and "red residue" may be identical.
Physiology might put more constraints on qualia, eventually ruling out all

other possibilities. But there may still be absent/alien qualia problems.

Hardin, C.L. 1997. Reinverting the spectrum. In (A. Byrne & D.R. Hilbert, eds) *_Readings on Color, Volume 1: The Philosophy of Color_*. MIT Press.

Harrison, B. 1967. On describing colors. *Inquiry* 10:38-52.

Harrison, B. 1973. *_Form and Content_*. Blackwell.
The inverted spectrum is impossible, due to asymmetries in color space.

Harvey, J. 1979. Systematic transposition of colours. *Australasian Journal of Philosophy* 57:211-19.

The inverted spectrum can be detected, if a single person experiences both.

Hatfield, G. 1992. Color perception and neural encoding: Does metameric matching entail a loss of information? *Philosophy of Science Association 1992*, 1:492-504.

Johnsen, B.C. 1986. The inverted spectrum. *Australasian Journal of Philosophy* 64:471-6.

Against Shoemaker: physical realizations do not give empirical conditions for qualia inversion. Nice.

Johnsen, B.C. 1993. The intelligibility of spectrum inversion. *Canadian Journal of Philosophy* 23:631-6.

Kirk, R. 1982. Goodbye to transposed qualia. *Proceedings of the Aristotelian Society* 82:33-44.

The possibility of an inverted spectrum w.r.t. dispositions implies the falsity of physicalism. But this rests on an implausible "slide-viewer" model of seeing, and is incoherent otherwise.

Levine, J. 1988. Absent and inverted qualia revisited. *Mind and Language* 3:271-87.

Inverted qualia, with respect to a functional account, are no more plausible than absent qualia (by analysis of thought experiments). Both lead to first-person skepticism about qualia.

Levine, J. 1991. Cool red. *Philosophical Psychology* 4:27-40.

Contra Hardin 1988: there's a "green residue" after coolness is subtracted, so inverted spectrum could still be possible. In any case, the impossibility of IS doesn't affect the explanatory gap for qualia, which is epistemic.

Lycan, W.G. 1973. Inverted spectrum. *Ratio* 15:315-9.

Inverted spectrum holding behavior constant is at least a coherent idea. Hook up brain in different ways, etc.

Lycan, W.G. 1993. Functionalism and recent spectrum inversions. Manuscript. Argues that qualia are intentional properties, and that inverted spectra, though conceivable, are metaphysically impossible, due to considerations about society and normality. Argues against Block's "inverted earth".

Nida-Rumelin, M. 1996. Pseudonormal vision: An actual case of qualia inversion? *Philosophical Studies* 82:145-57.

A fascinating note on the possibility of people with doubled colorblindness genes, thus inverting color processing; such people may actually exist.

Palmer, S. 1999. Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences*.

Rey, G. 1992. Sensational sentences reversed. *Philosophical Studies* 68:289-319.

Argues for a computational/sentential theory under which qualia are fixed by

functional organization. Argues against Block's 1990 inversion: qualia might slowly change back as associations fade. Memory isn't 100% reliable.

Shoemaker, S. 1975. Phenomenal similarity. *Critica* 7:3-37. Reprinted in *_Identity, Cause, and Mind_* (Cambridge University Press, 1984).

Maybe IS is ongoing, with memory changes. What is the logic of "appears"?

Shoemaker, S. 1982. The inverted spectrum. *Journal of Philosophy* 79:357-381. Reprinted in *_Identity, Cause, and Mind_* (Cambridge University Press, 1984).

All about the coherence and otherwise thereof. Uses switch in state for IS IS wrt behavior. Also claims that IS wrt function is possible as qualia are fixed by realizing state, not functional state. Bad assumption.

Shoemaker, S. 1996. Intersubjective/intrasubjective. In *_The First-Person Perspective and Other Essays_*. Cambridge University Press.

Taylor, D. 1966. The incommunicability of content. *Mind* 75:527-41. Inverted spectra thought-experiments show that experiential content is incommunicable. Accounts for the fact that attempts to describe such cases lead to contradiction (I'm seeing green & not seeing green).

Tye, M. 1993. Qualia, content, and the inverted spectrum. *Nous*. Argues that qualia are intentional properties, along the lines of "looks F to P". Handles inverted earth and related cases by taking the narrow intentional content. With remarks on the semantics of color terms.

1.7e Absent Qualia (Block, etc) [see also 1.3b]

Block, N. & Fodor, J.A. 1972. What psychological states are not. *Philosophical Review* 81:159-81.

As a criticism of functionalism. raises the possibility that realizations of any given functional account of mental states may lack qualia.

Block, N. 1980. Troubles with functionalism. In (N. Block, ed) *_Readings in the Philosophy of Psychology_*, Vol 1. Harvard University Press.

All kinds of absent qualia cases: homunculi-headed robots, the population of China, and so on. There is a prima facie doubt that such cases lack qualia, so there is a prima facie case against functionalism.

Bogen, J. 1981. Agony in the schools. *Canadian Journal of Philosophy* 11:1-21. It's OK for bizarre realizations to lack pain, as functionalism requires teleology as well as organization. With remarks on the relation between pain and "introspectible noxiousness".

Carleton, L. 1983. The population of China as one mind. *Philosophy Research Archives* 9:665-74.

Taking the personal stance, we should regard the Chinese nation as having qualia. A lack of qualia would make a functional difference.

Churchland, P.M. & Churchland, P.S. 1981. Functionalism, qualia and intentionality. *Philosophical Topics* 12:121-32. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).

Absent qualia are impossible. Also, qualia aren't essential to mental state, functional role is.

Cuda, T. 1985. Against neural chauvinism. *Philosophical Studies* 48:111-27. Replace neurons one by one with homunculi: what happens? Beliefs don't change, does consciousness fade? Very nice.

Elugardo, R. 1983. Functionalism, homunculi-heads and absent qualia. *Dialogue* 21:47-56.

If absent qualia are possible, then either qualia are inexplicable or species

chauvinism is true. Homunculi-heads could make similar arguments about us.

Elugardo, R. 1983. Functionalism and the absent qualia argument. *Canadian Journal of Philosophy* 13:161-80.

Hardcastle, V.G. 1996. Functionalism's response to the problem of absent qualia. *Journal of Consciousness Studies* 3:357-73.

Jacoby, H. 1990. Empirical functionalism and conceivability arguments. *Philosophical Psychology* 2:271-82.

Conceivability arguments are only a problem for empirical functionalism insofar as they are a problem for materialism in general. Very true.

Juhl, C.F. 1998. Conscious experience and the nontriviality principle. *Philosophical Studies* 91:91-101.

Levin, J. 1985. Functionalism and the argument from conceivability. *Canadian Journal of Philosophy Supplement* 11:85-104.

Argues that metaphysical conclusions can be drawn from conceivability arguments, but that absent qualia cases have not been clearly and distinctly conceived. The functionalist is better off than the identity theorist here.

Levine, J. 1988. Absent and inverted qualia revisited. *Mind and Language* 3:271-87.

IQ are no more plausible than AQ (by analysis of thought experiments and skepticism). So there's no reason to choose physicalist-functionalism over pure functionalism, as Shoemaker does. Nice.

Sayan, E. 1988. A closer look at the Chinese Nation argument. *Philosophy Research Archives* 13:129-36.

The Chinese Nation would require less people than Churchland & Churchland 1981 suggest, as we'd only need to handle a subset of all possible inputs.

Tye, M. 1993. Blindsight, the absent qualia hypothesis, and the mystery of consciousness. In (C. Hookway, ed) *Philosophy and the Cognitive Sciences*. Cambridge University Press.

Gives a thorough neurophysiological analysis of blindsight and related pathologies, and argues that these cannot be used to support the possibility of absent qualia. With remarks on the mystery of consciousness.

1.7f Introspection and Absent Qualia (Shoemaker) [see also 1.6d]

Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:291-315. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Absent qualia possible => qualia make no causal difference => no knowledge of qualia, therefore absent qualia are impossible. If qualia are introspectively accessible, they must be functional. An important argument.

Shoemaker, S. 1981. Absent qualia are impossible -- A reply to Block. *Philosophical Review* 90:581-99. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Reply to Block 1980. Distinguishes two AQ theses, and argues that if AQ are possible, then the problem for functionalism isn't due solely to qualia.

Averill, E.W. 1990. Functionalism, the absent qualia objection, and eliminativism. *Southern Journal of Philosophy* 28:449-67.

Defending Shoemaker's argument against Conee: immediate awareness and qualitative beliefs are the same. But maybe people **can't** tell whether they're having genuine or ersatz pain. Eliminativism is the best option.

Block, N. 1980. Are absent qualia impossible? *Philosophical Review* 89:257-74.

Reply to Shoemaker 1975. The possibility of absent qualia is compatible with a functional role for qualia, as qualia can make a causal difference that is independent of a given functional account.

Conee, E. 1985. The possibility of absent qualia. *Philosophical Review* 94:345-66.

Contra Shoemaker: qualia cause qualitative beliefs, which are affected by the absence of qualia, so we know about qualia even if AQ are possible.

Davis, L. 1982. Functionalism and absent qualia. *Philosophical Studies* 41:231-49.

Elucidating Shoemaker's argument: if absent qualia are possible, then the difference between real and ersatz pain makes no difference to belief, so qualia aren't introspectively accessible. A nice analysis.

Doore, G. 1981. Functionalism and absent qualia. *Australasian Journal of Philosophy* 59:387-402.

Qualia and qualitative beliefs are the same, so Shoemaker's argument fails. A numbness/pain inversion argument shows that pain isn't a functional state; it yields an introspectible difference without a functional difference.

Francescotti, R.M. 1994. Qualitative beliefs, wide content, and wide behavior. *Nous* 28:396-404.

Qualitative beliefs can supervene on behavioral dispositions even if absent/inverted qualia are possible. We just individuate belief contents and behavior widely, with wide content fixed to the qualia.

Hill, C.S. 1991. Introspection and the skeptic. In *Sensations: A Defense of Type Materialism*. Cambridge University Press.

Argues that the possibility of absent qualia is compatible with introspective knowledge. The fact that we have evidence of qualia isn't altered by the fact that we'd still think we had that evidence if we didn't have qualia.

White, N. 1985. Professor Shoemaker and the so-called 'qualia' of experience. *Philosophical Studies* 47:369-383.

Shoemaker's account leaves out experienced relations, such as experienced similarity. Experienced similarity is not the same as similarity between experiences. Being experienced is not an experienced feature.

1.7g Functionalism and Qualia, General

Brown, M. 1983. Functionalism and sensations. *Auslegung* 10:218-28.

Various comments on functionalism's troubles with qualia, including absent and inverted qualia. Analogis with biology and information theory.

Chalmers, D.J. 1995. Absent qualia, fading qualia, dancing qualia. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh. Press.

Argues that absent qualia and inverted qualia are empirically impossible (though logically possible), using neural-replacement thought-experiments. So functional organization fully determines conscious experience.

Cole, D.J. 1990. Functionalism and inverted spectra. *Synthese* 82:207-22.

Acquired spectrum inversion doesn't refute functionalism, if qualia revert after behavioral adaptation. With empirical evidence.

Dumpleton, S. 1988. Sensation and function. *Australasian Journal of Philosophy* 66:376-89.

Eshelman, L.J. 1977. Functionalism, sensations, and materialism. *Canadian Journal of Philosophy* 7:255-74.

Graham, G. & Stephens, G. 1985. Are qualia a pain in the neck for functionalists? *American Philosophical Quarterly* 22:73-80.

Pain-qualia are in the body, not the mind, and so aren't part of psychology.

Graham, G. & Stephens, G. 1987. Minding your P's and Q's: Pain and sensible qualities. *Nous* 21:395-405.

Greenberg, W. 1998. On Chalmers' "principle of organizational invariance" and his "dancing qualia" and "fading qualia" thought experiments. *Journal of Consciousness Studies* 5:53-58.

Hill, C.S. 1991. The failings of functionalism. In *Sensations: A Defense of Type Materialism*. Cambridge University Press.

Gives a number of arguments against both analytic functionalism and psychofunctionalism: arguments from absent qualia, absent functional role, epistemology, semantics, and heterogeneity of functional roles.

Horgan, T. 1984. Functionalism, qualia, and the inverted spectrum. *Philosophy and Phenomenological Research*, 44:453-69.

Argues that non-phenomenal mental events are functional, while qualia are low-level physiological.

Jarrett, G. 1996. Analyzing mental demonstratives. *Philosophical Studies* 84:49-62.

Lycan, W.G. 1981. Form, function and feel. *Journal of Philosophy* 78:24-50.

Accuses Block of a perspective error. Functionalism can handle a lot, if it's multi-levelled.

Lycan, W.G. 1987. Homunctionalism and qualia. In *Consciousness*. MIT Press.

Various stuff, mostly against absent qualia arguments.

Moor, J.H. 1988. Testing robots for qualia. In (H. Otto & J. Tuedio, eds) *Perspectives on Mind*. Kluwer.

Behavioral evidence for qualia is always indirect. And you can't check by replacing own neurons by chips, as you'll still believe you have qualia if you're functionally identical. Posit robot qualia as explanatory construct?

Nemirow, L. 1979. Functionalism and the subjective quality of experience. Dissertation, Stanford University.

Rey, G. 1994. Wittgenstein, computationalism, and qualia. In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*.

Holder-Pichler-Tempsky.

Computational functionalism about qualia is compatible with Wittgenstein's views. It makes sense of the points about "dividing through" my private objects, for example. With remarks on spectrum inversions.

Seager, W.E. 1983. Functionalism, qualia and causation. *Mind* 92:174-88.

Functionalism can't explain the causal role of qualia by identifying them with functional states (circularity) or physical realizations (chauvinism). Which leaves property dualism, epiphenomenalism, or eliminativism for qualia.

Shoemaker, S. 1994. The first-person perspective. *Proceedings and Addresses of the American Philosophical Association* 68:7-22.

Against drawing strong conclusions from first-person imaginings. Considers Searle's silicon-replacement scenario: we might infer that perception isn't veridical, that there's another mind about, or even another body.

van Gulick, R. 1988. Qualia, functional equivalence and computation. In (H. Otto & J. Tuedio, eds) *Perspectives on Mind*. Kluwer.

Commentary on Moor 1988. Systems that differ in qualitative properties will

likely differ in functional organization.

van Heuveln, B., Dietrich, E. & Oshima, M. 1998. Let's dance! The equivocation in Chalmers' dancing qualia argument. Minds and Machines.

White, S. 1986. Curse of the qualia. Synthese 68:333-68.
Criticism of "physicalist-functionalism", where functional organization doesn't completely determine qualia (e.g. Shoemaker/Block). The only tenable options are pure functionalism or transcendental dualism. Nice.

White, S. 1989. Transcendentalism and its discontents. Philosophical Topics 17:231-61.
Taking transcendental dualism seriously. Privileged access provides strong arguments against objective theories, but it turns out that transcendentalism can't explain it any better, so maybe embrace objective theories after all.

Wright, E. 1995. More qualia trouble for functionalism: The Smythies TV-hood analogy. Synthese 97:365-82.

Zuboff, A. 1994. What is a mind? Midwest Studies in Philosophy 19:183-205.
Replacing a brain chunk while preserving causal role must preserve experience;

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 2: Mental Content [869]

=====

- 2.1 Propositional Attitudes and Folk Psychology [285]
 - 2.1a The Language of Thought (Fodor) [54]
 - 2.1b The Intentional Stance (Dennett) [40]
 - 2.1c Eliminativism (Churchlands, etc) [57]
 - 2.1d Propositional Attitudes, General [42]
 - 2.1e The Nature of Folk Psychology [38]
 - 2.1f The Simulation Theory [54]
- 2.2 Internalism and Externalism [278]
 - 2.2a Is Content in the Head? (Putnam, Burge) [38]
 - 2.2b Externalism and Psychological Explanation [41]
 - 2.2c Externalism and Mental Causation [28]
 - 2.2d Externalism and the Theory of Vision [16]
 - 2.2e Externalism and Computation [11]
 - 2.2f Externalism and Self-Knowledge [72]
 - 2.2g The Status of Narrow Content [43]
 - 2.2h Miscellaneous [29]
- 2.3 Theories of Content [144]
 - 2.3a Information-Based Accounts (Dretske, etc) [29]
 - 2.3b Asymmetric Dependence (Fodor) [20]
 - 2.3c Causal Accounts, General [17]
 - 2.3d Teleological Approaches (Millikan, etc) [48]
 - 2.3e Conceptual Role Approaches [19]
 - 2.3f Theories of Content, Misc [11]
- 2.4 Representation, General [25]
- 2.5 The Explanatory Role of Content (Dretske, etc) [22]
- 2.6 Concepts [46]
- 2.7 Meaning Holism [26]
- 2.8 Mental Content, Misc [43]

2.1 The Status of Propositional Psychology

2.1a The Language of Thought (Fodor)

Fodor, J.A. 1975. *The Language of Thought*. Harvard University Press.
Argues that thought involves computation upon representations, and that these are structured as sentences in a mental language. With linguistic and psychological evidence, and arguments that the mental language is innate.

Fodor, J.A. 1987. Why there still has to be a language of thought. In *Psychosemantics*. MIT Press.
Because it fits explanatory methodology, it coheres with the usual ontology of psychological processes, and it explains systematicity.

Fodor, J.A. 1978. Propositional attitudes. *Monist* 61:501-23. Reprinted in *RePresentations* (MIT Press, 1980).
About what PA's are, and why they're at the foundations of thought.

Fodor, J. 2001. Language, thought and compositionality. *Mind and Language* 16:1-15.

Abbott, B. 1995. Natural language and thought: Thinking in English. *Behavior and Philosophy* 23:49-55.

Bonjour, L. 1991. Is thought a symbolic process? *Synthese* 89:331-52.
Argues that symbol processing can't account for the intrinsically contentful nature of thought: using a symbol doesn't give understanding of its content. With defense against arguments from twin earth and conceptual-role semantics.

Braddon-Mitchell, D. & Fitzpatrick, J. 1990. Explanation and the language of thought. *Synthese* 83:3-29.

No need to postulate LOT: diachronic explanation is as good as synchronic, and high-level laws can exist without high-level causal connections.

Clapin, H. 1997. Problems with principle P. *Pacific Philosophical Quarterly* 78:261-??.

Clark, A. 1988. Thoughts, sentences and cognitive science. *Philosophical Psychology* 1:263-78.

Crane, T. 1990. The language of thought: No syntax without semantics. *Mind and Language* 5:187-213.

Davies, M. 1992. Aunty's own argument for the language of thought. In (J. Ezquerro & J. Larrazabal, eds) *Cognition, Semantics and Philosophy*. Kluwer.

Dennett, D.C. 1977. A cure for the common code. *Mind*. Reprinted in *Brainstorms* (MIT Press, 1978).

Review of Fodor's LOT. Fodor's view is too strong: function, not structure, is criterial for content. The structure of a predictive theory need not be directly reflected in inner processing.

Dennett, D.C. 1975. Brain writing and mind reading. *Minnesota Studies in the Philosophy of Science* 7:403-15. Reprinted in *Brainstorms* (MIT Press, 1978).

On the explicit representation of belief: criteria, plausibility, and relationship to verbal reports and conscious judgments.

Dennett, D.C. 1990. Granny's campaign for safe science. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

A general treatment of Fodor, identifying him as arch-conservative mentalist.

DeWitt, R. 1995. Vagueness, semantics, and the language of thought. *Psyche* 1.

Dunlop, G. 1990. Conceptual dependency as the language of thought. *Synthese* 82:275-96.

Relates Schank's conceptual dependency to Fodor's LOT.

Egan, M.F. 1991. Propositional attitudes and the language of thought. *Canadian Journal of Philosophy* 21:379-88.

Contra two of Fodor's arguments for LOT. Complex causes need not have LOT constituency structure; and evidence from psychological theory falls short.

Field, H. 1978. Mental representation. *Erkenntnis* 13:9-18. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Analyzes belief into a relation between a person and an internal sentence, along with a semantic relation between that sentence and e.g. a proposition. With arguments against functionalist analyses, and against propositions.

Garson, J.W. 1997. Syntax in a dynamic brain. *Synthese* 110:343-55.

There are no good arguments for LOT of the form "The brain needs to do X, and X entails LOT". Considers X = concatenation, logical form, tracking, combinatorial encoding. Either LOT is weakened deeply or is unnecessary.

Garfield, J. 2000. Thought as language: A metaphor too far. *Protosociology* 14:85-101.

Gauker, C. 1995. *Thinking Out Loud: An Essay on the Relation between Thought and Language*. Princeton University Press.

Harman, G. 1973. *Thought*. Princeton University Press.

Harman, G. 1975. Language, thought, and communication. *Minnesota Studies in the Philosophy of Science* 7:270-298.

Argues that the primary role of language is in thought rather than in communication, and the language of thought incorporates natural language.

Harman, G. 1977. How to use propositions. *American Philosophical Quarterly*.

Harman, G. 1978. Is there mental representation? *Minnesota Studies in the Philosophy of Science* 9.

Hauser, L. 1995. Natural language and thought: Doing without mentalese. *Behavior and Philosophy* 23:41-47.

Heil, J. 1981. Does cognitive psychology rest on a mistake? *Mind* 90:321-42.
LOT confuses processes with descriptions of processes. Also, symbols cannot denote solely in virtue of structure, so must rely on human interpretation.

Kaye, L.J. 1994. The computational account of belief. *Erkenntnis* 40:137-53.

Kaye, L.J. 1995. The languages of thought. *Philosophy of Science* 62:92-110.

Knowles, J. 1998. The language of thought and natural language understanding. *Analysis* 58:264-272.

Loar, B. 1982. Must beliefs be sentences? *Philosophy of Science Association*.

Lycan, W.G. 1982. Toward a homuncular theory of believing. *Cognition and Brain Theory* 4:139-59.

Defends sententialism of the homuncular variety: little modules all the way in. Lots of pro-belief arguments.

Lycan, W.G. 1990. Mental content in linguistic form. *Philosophical Studies* 58:147-54.

Distinguishes varieties of Sententialism, reasonable vs. mad-dog.

Lycan, W.G. 1993. A deductive argument for the representational theory of thinking. *Mind and Language* 8:404-22.

Argues from the unboundedness of thinking and the need for a finite stock of elements to something like a language of thought. With remarks on connectionism and instrumentalism, and a reply by Stalnaker.

Laurence, S. & Margolis, E. 1997. Regress arguments for the language of thought. *Analysis* 57:60-66.

Marras, A. 1987. The weak and the strong representational theory of mind: Stich's interpretation of Fodor. *Dialogue* 26:349-55.

Matthews, R.J. 1989. The alleged evidence for representationalism. In (S. Silvers, ed) *_Rerepresentation_*. Kluwer.

Argues that contrary to some claims, cognitive psychology does not provide much support for a computational/representational theory of propositional attitudes. Specifically considers research in psycholinguistics and vision.

Matthews, R.J. 1991. Is there vindication through representationalism? In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell.

Fodor's theory can't deal with inexplicit attitudes: the core/derivative distinction is untenable. But we can make sense of intentional causation and psychological explanation without explicit representation.

Millikan, R.G. 1993. On mentalese orthography. In (B. Dahlbom, ed) *_Dennett and his Critics_*. Blackwell.

On some problems typing tokens in the language of thought. There's no principled distinction between type-identical tokens and type-distinct tokens

with an identity judgment. With interesting remarks on co-identification.

Pessin, A. 1995. Mentalese syntax: Between a rock and two hard places. *Philosophical Studies* 78:33-53.

Argues that there is no good way to individuate syntactic types in Mentalese. Neural typing, causal typing, and semantic typing all fail.

Rantala, V. & Vaden, T. 1997. Minds as connoting systems: Logic and the language of thought. *Erkenntnis* 46:315-334.

Rey, G. 1995. A not "merely empirical" argument for the language of thought. *Philosophical Perspectives* 9:201-22.

Schiffer, S. 1991. Does Mentalese have a compositional semantics? In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell. Argues that the language of thought need not have a compositional semantics; productivity and systematicity can be explained without it.

Schiffer, S. 1994. The language-of-thought relation and its implications. *Philosophical Studies* 76:263-85.

Schwartz, G. 1996. Symbols and thought. *Synthese* 106:399-407.

Sher, G. 1975. Sentences in the brain. *Philosophy and Phenomenological Research* 36:94-99.

On Danto's suggestion that beliefs are like sentences. Conventionality poses problems, as does differentiating between different sorts of attitudes.

Stalnaker, R.C. 1990. Mental content and linguistic form. *Philosophical Studies* 58:129-46.

Sterelny, K. 1983. Mental representation: What language is Brainese? *Philosophical Studies*, 43:365-82.

Motivates LOT and defends it against various objections: e.g. tacit belief, identity conditions, infinite regress, and semantic nativism.

Stich, S.P. 1978. Beliefs and subdoxastic states. *Philosophy of Science* 45:499-518.

Teng, N.Y. 1999. The language of thought and the embodied nature of language use. *Philosophical Studies* 94:237-251.

Tienson, J. 1990. Is this any way to be a realist? *Philosophical Psychology*.

Warmbrod, K. 1989. Beliefs and sentences in the head. *Synthese* 2:201-30.

Weller, C. 1997. Bonjour and mentalese. *Synthese* 113:251-63.

Yagisawa, T. 1994. Thinking in neurons: Comments on Stephen Schiffer's "The language-of-thought relation and its implications". *Philosophical Studies* 76:287-96.

2.1b The Intentional Stance (Dennett)

Dennett, D.C. 1978. *Brainstorms*. MIT Press.

Dennett, D.C. 1971. Intentional systems. *Journal of Philosophy* 68:87-106
Reprinted in *Brainstorms* (MIT Press, 1978).

Can view systems from physical stance, design stance, or intentional stance. Beliefs/desires are attributed under the intentional stance, with help from certain idealized norms of rationality and accuracy licensed by evolution.

Dennett, D.C. 1981. Making sense of ourselves. *Philosophical Topics* 12:63-81.

Reprinted in The Intentional Stance (MIT Press, 1987).

Reply to Stich 1981. Irrationality is misdesign (take design stance). Etc.

Dennett, D.C. 1987. The Intentional Stance. MIT Press.

Beliefs/desires are useful predictive attributions. This isn't inconsistent with a certain degree of realism (abstracta/illata distinction).

Dennett, D.C. 1988. Precis of The Intentional Stance. Behavioral and Brain Sciences.

TIS, with commentaries and replies.

Dennett, D.C. 1990. The interpretation of texts, people and other artifacts. Philosophy and Phenomenological Research (Supplement) 50.

Mental states are underdetermined: like interpreting a text, or finding an object's function. Even adaptationist teleology gives no fact of the matter.

Dennett, D.C. 1991. Real patterns. Journal of Philosophy 88:27-51.

Proposition attitudes have the ontological status of a noisy pattern that helps make sense of behavior. This degree of realism falls on a scale: Fodor > Davidson > Dennett > Rorty > Churchland.

Baker, L.R. 1987. Instrumentalism: Back from the brink? In Saving Belief. Princeton University Press.

Dennett vacillates between stance-dependence, -independence; e.g. on rationality, design features. Instrumentalism can't be rescued.

Baker, L.R. 1989. Instrumental intentionality. Philosophy of Science 56:303-16.

Bechtel, W. 1985. Realism, instrumentalism, and the intentional stance. Cognitive Science 9:265-92.

Dennett should be a realist, of the relative-to-environment variety.

Byrne, A. 1998. Interpretivism. European Review of Philosophy 3.

Cam, P. 1984. Dennett on intelligent storage. Philosophy and Phenomenological Research 45:247-62.

Clark, A. 1990. Belief, opinion and consciousness. Philosophical Psychology. Argues contra Dennett and Smolensky that language is fundamental, not just an add-on.

Cohen, B. 1995. Patterns lost: Indeterminism and Dennett's realism about beliefs. Pacific Philosophical Quarterly 76:17-31.

Cummins, R. 1981. What can be learned from Brainstorms? Philosophical Topics 12:83-92.

Questioning Dennett on the bridge between intentional characterization and functional characterization. Arguing for the importance of context.

Davies, D. 1995. Dennett's stance on intentional realism. Southern Journal of Philosophy 33:299-312.

Fodor, J.A. 1981. Three cheers for propositional attitudes. In Representations. MIT Press.

Dennett's rationality/intentional idealization assumptions should not be viewed as Platonic but epistemic. PA's are real and play real roles.

Fodor, J.A. & LePore, E. 1993. Is intentional ascription intrinsically normative? In (B. Dahlbom, ed) Dennett and His Critics. Blackwell.

Against "interpretivism" about intentionality: projectivism is hopeless, and Dennett's arguments for normativism (via charity and evolution) go wrong or beg the question.

Foss, J. 1994. On the evolution of intentionality as seen from the intentional stance. *Inquiry* 37:287-310.

Gauker, C. 1988. Objective interpretationism. *Pacific Philosophical Quarterly* 69:136-51.

Haugeland, J. 1993. Pattern and being. In (B. Dahlbom, ed) *_Dennett and His Critics_*. Blackwell.

Heitner, R. 2000. Is design relative or real? Dennett on intentional relativism and physical realism. *Minds and Machines* 10:267-83.

Hornsby, J. 1992. Physics, biology, and common-sense psychology. In (D. Charles & K. Lennon, eds) *_Reduction, Explanation and Realism_*. Oxford University Press.

Kukla, R. 2000. How to get an interpretivist committed. *Protosociology* 14:180-221.

Lyons, W. 1990. Intentionality and modern philosophical psychology, I. The modern reduction of intentionality. *Philosophical Psychology* 3:247-69.

McLaughlin, B. & O'Leary-Hawthorne, J. 1995. Dennett's logical behaviorism. *Philosophical Topics* 22:189-258.

McLaughlin, B. 2000. Why intentional systems theory cannot reconcile physicalism with realism about belief and desire. *Protosociology* 14:145-157.

McCulloch, G. 1990. Dennett's little grains of salt. *Philosophical Quarterly* 40:1-12.

Dennett must be one of: realist, eliminativist, instrumentalist.

Narayanan, A. 1996. The intentional stance and the imitation game. In (P. Millican & A. Clark, eds) *_Machines and Thought_*. Oxford University Press.

Nelkin, N. 1993. Patterns. *Mind and Language* 9:56-87.

Dennett's instrumentalism can't explain the acquisition of intentional concepts. Proposition attitudes are directly introspectible entities, although still theoretical and still patterns.

Price, H. 1995. Psychology in perspective. In (M. Michael & J. O'Leary-Hawthorne, eds) *_Philosophy in Mind_*. Kluwer.

Radner, D. & Radner, M. 1995. Cognition, natural selection, and the intentional stance. *International Studies in the Philosophy of Science* 9:109-19.

Richard, M. 1995. What isn't a belief? *Philosophical Topics* 22:291-318.

Richardson, R.C. 1980. Intentional realism or intentional instrumentalism? *Cognition and Brain Theory* 3:125-35.

Sharpe, R. 1989. Dennett's journey towards panpsychism. *Inquiry* 32:233-40.

Slors, M. 1996. Why Dennett cannot explain what it is to adopt the intentional stance. *Philosophical Quarterly* 46:93-98.

Stich, S.P. 1980. Headaches. *Philosophical Books* 21:65-73.
Critical review of *_Brainstorms_*, with response.

Stich, S.P. 1981. Dennett on intentional systems. *Philosophical Topics* 12:39-62. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).
Dennett has problems with rationality, realism, etc. Hard line/soft line:

either intentional stance is too close to FP or too far away.

Webb, S. 1994. Witnessed behavior and Dennett's intentional stance. *Philosophical Topics* 22:457-70.

Yu, P. & Fuller, G. 1986. A critique of Dennett. *Synthese* 66:453-76.
Very thorough account of the evolution of Dennett's views. Elucidates abstracta/illata, criticizes intentional subpersonal psychology.

2.1c Eliminativism (Churchlands) [see also 4.3c]

Churchland, P.S. 1980. Language, thought, and information processing. *Nous* 14:147-70.
Sentential processing is out. Against Harman's mental English and Fodor's Mentalese. Arguments from learning, evolution, neuroscience, mental images.

Churchland, P.M. 1981. Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78:67-90. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).
Eliminate beliefs/desires, remnants of a stagnant folk theory.

Churchland, P.M. & Churchland, P.S. 1983. Stalking the wild epistemic engine. *Nous* 17:5-20. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).
How to dethrone language and still handle content.

Churchland, P.M. 1985. On the speculative nature of our self-conception. *Canadian Journal of Philosophy Supplement* 11:157-173.
Reply to Foss 1985: EM is plausible, though certainly not applicable everywhere -- e.g. sensations will be reduced, not eliminated.

Churchland, P.M. 1989. *_A Neurocomputational Perspective: The Nature of Mind and the Structure of Science_*. MIT Press.
14 glimpses of the neurophilosophical golden age.

Churchland, P.M. 1993. Theory, taxonomy, and methodology: A reply to Haldane's "Understanding folk". *Proceedings of the Aristotelian Society* 67:313-19.
Reply to Haldane 1988. Even observations can be reconceived. With remarks perceptual plasticity and propositions, and a rejoinder by Haldane.

Churchland, P.M. 1993. Evaluating our self-conception. *Mind and Language* 8:211-22.
It's "bad faith" to accept modern epistemology but to deny the possibility of eliminativism. On various objections: "functional kinds", "self-defeating", "what could falsify it?", "different purposes", "no alternatives".

Baker, L.R. 1987. The threat of cognitive suicide. In *_Saving Belief_*. Princeton University Press.
Elaborating the paradoxes of disbelieving in belief. Rational acceptability, assertion, and truth are all at risk.

Baker, L.R. 1988. Cognitive suicide. In (R. Grimm & D. Merrill, eds) *_Contents of Thought_*. University of Arizona Press.
Eliminativism is pragmatically incoherent, as it implies that language isn't meaningful and that the thesis isn't formulable. Folk psychology needn't be scientifically reduced to be true. With comments by Chastain, and reply.

Bertolet, R. 1994. Saving eliminativism. *Philosophical Psychology* 7:87-100.
Against Baker's cognitive-suicide arguments against eliminativism. We don't know what a replacement theory will look like, but that doesn't show that none is forthcoming.

- Bickle, J. 1992. Revisionary physicalism. *Biology and Philosophy* 7:411-30.
Argues for a revisionary reduction of the propositional attitudes, rather than elimination or smooth reduction. Sentential aspects will go, but coarse-grained functional profiles and content will remain.
- Blunt, P.K. 1992. A defense of folk psychology. *International Philosophical Quarterly* 32:487-98.
- Boghossian, P. 1990. The status of content. *Philosophical Review* 99:157-84.
Irrealism about mental content (and therefore truth-conditions) can't be made sense of. An error thesis presupposes factual truth-conditions, and a non-factualist thesis presupposes a non-deflationary theory of truth.
- Boghossian, P. 1991. The status of content revisited. *Pacific Philosophical Quarterly* 71:264-78.
Reply to Devitt 1990.
- Chater, N. & Oaksford, M. 1996. The falsity of folk theories: Implications for psychology and philosophy. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.
- Clark, A. 1996. Dealing in futures: Folk psychology and the role of representations in cognitive science. In (R. McCauley, ed) *The Churchlands and their Critics*. Blackwell.
- Cling, A. 1989. Eliminative materialism and self-referential inconsistency. *Philosophical Studies* 56:53-75.
Unbelief in belief is not incoherent. Argues with Baker.
- Cling, A. 1990. Disappearance and knowledge. *Philosophy of Science* 57:226-47.
- Cling, A. 1991. The empirical virtues of belief. *Philosophical Psychology* 4:303-23.
Cognitive states like belief are necessary to explain the dependence of behavior on perceptual features of the environment. Informational states alone are not enough, as they can't explain selective response to features.
- Devitt, M. 1990. Transcendentalism about content. *Pacific Philosophical Quarterly* 71:247-63.
Against Boghossian's critique: the eliminativism will express her claim in a new framework, so appeals to truth beg the question. With a response.
- Devitt, M. & Rey, G. 1991. Transcending transcendentalism. *Pacific Philosophical Quarterly* 72:87-100.
Rejoinder to Boghossian 1990.
- Foss, J.E. 1985. A materialist's misgivings about eliminative materialism. *Canadian Journal of Philosophy Supplement* 11:105-33.
EM needs much more evidence before being so gung ho.
- Graham, G. & Horgan, T. 1992. Southern fundamentalism and the end of philosophy. In (E. Villanueva, ed) *Truth and Rationality*. Ridgeview.
- Greenwood, J.D. 1991. Reasons to believe. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.
Argues that folk psychological states exist, even if they aren't useful as causal explanation. We have independent reason to believe in them, e.g. from self-knowledge. They're useful in social psychology, too.
- Greenwood, J.D. 1992. Against eliminative materialism: from folk psychology to Volkerpsychologie. *Philosophical Psychology* 5:349-68.

Haldane, J. 1988. Understanding folk. Aristotelian Society Supplement 62:222-46.

Argues that folk psychology is not a theory, and that psychological knowledge is a pre-theoretical given. With remarks on laws, the prediction of behavior, and neuroscience.

Hannan, B. 1990. 'Non-scientific realism' about propositional attitudes as a response to eliminativist arguments. Behavior and Philosophy 18:21-31.

Hannan, B. 1993. Don't stop believing: the case against eliminative materialism. Mind and Language 8:165-179.

A bundle of arguments against eliminativism, e.g. from incoherence, the lack of alternatives, and against the folk-theory-theory. With commentary.

Horgan, T. & Woodward, J. 1985. Folk psychology is here to stay. Philosophical Review 94:197-225. Reprinted in (W. Lycan, ed) _Mind and Cognition_ (Blackwell, 1990).

Defending folk psychology against the arguments of Churchland and Stich: e.g. incompleteness, stagnation, irreducibility, dual-control, modularity, and unfalsifiability. Even with no neat reduction, folk psychology may be OK.

Horgan, T. & Graham, G. 1990. In defense of Southern Fundamentalism. Philosophical Studies 62:107-134.

FP is almost certainly true, irrespective of scientific absorbability or the language of thought. FP's commitments are austere, and mostly behavioral. Arguments from semantic competence and conceptual conservatism.

Horgan, T. 1993. The austere ideology of folk psychology. Mind and Language. Argues that FP is not committed to much. The austere conception is supported by intuitions, conservatism, and the inconceivability of dropping it. Responds to phlogiston objections: they are not analogous.

Horst, S. 1995. Eliminativism and the ambiguity of 'belief'. Synthese 104:123-45.

Clarifies different senses of "theoretical" and "belief". Some beliefs are relevantly theoretical (dispositional, infra-conscious, unconscious ones), but conscious occurrent beliefs are not, and so can't be eliminated.

Jackson, F. & Pettit, P. 1990. In defense of folk psychology. Philosophical Studies 59:31-54.

FP holds that beliefs/desires play a certain functional role, and it's almost certain that objects playing that role exist, so FP is fine, whether or not propositional attitudes are good scientific entities.

Jacoby, H. 1985. Eliminativism, meaning and qualitative states. Philosophical Studies.

Even if nothing satisfies all or most common-sense properties of mental terms, reference can still be fixed under a Putnam style theory of meaning. (More about qualia than about intentional states.)

Kitcher, P.S. 1984. In defense of intentional psychology. Journal of Philosophy 81:89-106.

The Churchlands underestimate the resources of intentional psychology.

Lahav, R. 1992. The amazing predictive power of folk psychology. Australasian Journal of Philosophy 70:99-105.

Melnyk, A. 1996. Testament of a recovering eliminativist. Philosophy of Science 63:S185-93.

O'Brien, G. 1987. Eliminative materialism and our psychological self-knowledge. Philosophical Studies 52:49-70.

Uses empirical evidence to argue that there is prelinguistic awareness, so

nominalistic arguments for eliminativism fail. And some awareness is innate, so we can't reconceive things in less than evolutionary time.

Ramsey, W. 1990. Where does the self-refutation objection take us? *Inquiry* 33:453-65.

The self-refutation objection reduces to other standard objections: counterexample, promissory note or reductio.

Ramsey, W., Stich, S.P. & Garon, J. 1991. Connectionism, eliminativism, and the future of folk psychology. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

If connectionism is true, then eliminativism is true, as you can't isolate the causal role of individual beliefs in a connectionist system.

Reppert, V. 1991. Ramsey on eliminativism and self-refutation. *Inquiry* 34:499-508.

Response to Ramsey 1990: If there are no beliefs and so no assertions, there is no identifiable propositional content, and truth and knowledge are out. Eliminativism is pragmatically self-refuting.

Reppert, V. 1992. Eliminative materialism, cognitive suicide, and begging the question. *Metaphilosophy* 23:378-92.

A careful analysis of whether self-refutation arguments against eliminativism beg the question by supposing that assertion requires belief. An account of what it is to beg the question, and a comparison to arguments about vitalism.

Resnick, P. 1994. Intentionality is phlogiston. In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

Richards, G. 1996. On the necessary survival of folk psychology. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Robinson, W.S. 1985. Toward eliminating Churchland's eliminationism. *Philosophical Topics* 13:60-67.

There's no reason to abandon FP, even if it doesn't reduce.

Rosenberg, A. 1991. How is eliminative materialism possible? In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.

Explaining how singular causal claims based on FP may be true even if FP is false; by analogy with phlogiston, and also because of near-vacuousness. EM isn't incoherent, as we can use a non-intentional replacement for belief.

Rosenberg, A. 1999. Naturalistic epistemology for eliminative materialists. *Philosophy and Phenomenological Research* 59:335-358.

Saidel, E. 1992. What price neurophilosophy? *Philosophy of Science Association* 1:461-68.

Folk psychology is compatible with neuroscientific models, but it need not smoothly reduce to neuroscience to have an important role.

Schouten, M.K.D. & de Jong, H.L. 1998. Defusing eliminative materialism: Reference and revision. *Philosophical Psychology* 11:489-509.

Schwartz, J. 1991. Reduction, elimination, and the mental. *Philosophy of Science* 58:203-20.

Stich, S.P. 1991. Do true believers exist? *Aristotelian Society Supplement* 65:229-44.

Eliminativism may have no determinate truth-conditions: if folk psychology is a poor theory, the question of whether or not "belief" refers may be empty.

Stich, S.P. 1992. What is a theory of mental representation? *Mind* 101:243-61.

Philosophical analysis isn't sufficient to understand intentional concepts; real cognitive science is required, with conceptual revision. The truth of eliminativism will be relative to the theory of reference that we choose.

Stich, S.P. 1996. Deconstructing the mind. In Deconstructing the Mind. Oxford University Press, 1996.

Taylor, K.A. 1994. How not to refute eliminative materialism. *Philosophical Psychology* 7:101-125.

Against transcendental arguments against eliminativism. These fail on their own terms, and even if successful they would not establish causal/explanatory relevance for the attitudes, which is the real key for folk psychology.

Tomberlin, J. 1994. Whither Southern Fundamentalism? In (E. Villanueva, ed) Truth and Rationality, Ridgeview.

Trout, J.D. 1991. Belief attribution in science: Folk psychology under theoretical stress. *Synthese* 87:379-400.

Wright, C. 1996. Can there be a rationally compelling argument for anti-realism about ordinary ("folk") psychology? In (E. Villanueva, ed) Content. Ridgeview.

2.1d Propositional Attitudes, General

Audi, R. 1994. Dispositional beliefs and dispositions to believe. *Nous* 28:419-34.

Baker, L.R. 1987. Saving Belief. Princeton University Press.
Beliefs are OK, despite no physicalist reduction of content.

Baker, L.R. 1993. What beliefs are not. In (S. Wagner & R. Warner, eds) Naturalism: A Critical Appraisal. University of Notre Dame Press.
Against beliefs construed as physically realized internal causes of behavior: syntax of these states can't be determinate, and their explanatory role wrt causation leads to a circle. Belief is irreducible.

Baker, L.R. 1994. Attitudes as nonentities. *Philosophical Studies* 76:175-203.

Balaguer, M. 1998. Attitudes without propositions. *Philosophy and phenomenological research* 58:805-26.

Bennett, J. 1991. Analysis without noise. In (R. Bogdan, ed) Mind and Common Sense. Cambridge University Press.

Remarks on the conceptual analysis of belief/desire attribution. On the roles of causation, inner-route explanations, belief-desire-action triangles, teleology, unity, the presumption of simplicity, and evolution.

Bennett, J. 1991. Folk-psychological explanations. In (J. Greenwood, ed) The Future of Folk Psychology. Cambridge University Press.

On requirements for belief/desire explanations: input/output patterns, the unity condition (i.e. no single associated mechanism), and teleological bases for generalizations, e.g. through evolution or educability.

Ben-Yami, H. 1997. Against characterizing mental states as propositional attitudes. *Philosophical Quarterly* 186:84-89.

Butler, K. 1992. The physiology of desire. *Journal of Mind and Behavior* 13:69-88.

Argues that desire will smoothly reduce to a neurophysiological kind.

Clark, A. 1991. Radical ascent. *Aristotelian Society Supplement* 65:211-27.

The conditions on being a believer are mostly behavioral; to claim otherwise is to fall into a "modularity trap". A counterfactual account of mental causation is enough. With a defense of mentality for giant look-up tables.

Clark, A. 1994. Beliefs and desires incorporated. *Journal of Philosophy* 91:404-25.

Cohen, L.J. 1996. Does belief exist? In (A. Clark & P. Millican, eds) *_Connectionism, Concepts, and Folk Psychology_*. Oxford University Press.

Crimmins, M. 1992. Tacitness and virtual beliefs. *Mind and Language* 7:240-63.

Davies, D. 1995. Davidson, indeterminacy, and measurement. *Acta Analytica* 10:37-56.

Davies, D. 1998. On gauging attitudes. *Philosophical Studies* 90:129-54.

Egan, M.F. 1989. What's wrong with the Syntactic Theory of Mind. *Philosophy of Science* 56:664-74.

Stich is confused about type-token, syntax/content, etc.

Fodor, J.A. 1986. Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind* 94:76-100. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

A taxonomy of positions on the representation of propositional attitudes: dividing up via questions about realism, functionalism, monadicity, and truth-conditions. With arguments for structured representations.

Frankish, K. 1998. A matter of opinion. *Philosophical Psychology* 11:423-442.

Garfield, J. 1988. *_Belief in Psychology: A Study in the Ontology of Mind_*. MIT Press.

Graham, G. & Horgan, T. 1988. How to be realistic about folk psychology. *Philosophical Psychology* 1.

Jacquette, D. 1990. Intentionality and Stich's theory of brain sentence syntax. *Philosophical Quarterly*, 40:169-82.

Things are only syntactic (in SS's sense) in virtue of intentionality. True.

Lycan, W.G. 1986. Tacit belief. In (R. Bogdan, ed) *_Belief: Form, Content, and Function_*. Oxford University Press.

Maloney, J.C. 1990. It's hard to believe. *Mind and Language* 5:122-48.

Manfredi, P.A. 1993. Tacit beliefs and other doxastic attitudes. *Philosophia*. Argues that there are no tacit beliefs: dispositions to believe can do all the explanatory work at lower cost. With some remarks on subdoxastic states, and the difference between belief and opinion.

Matthews, R.J. 1994. The measure of mind. *Mind* 103:131-46.
A theory of propositional attitude ascription as like numerical measurement.

Millikan, R.G. 1986. Thoughts without laws: Cognitive science with content. *Philosophical Review* 95:47-80.

Folk psychology isn't a theory about laws, but about proper functions.
desires are identified by proper functions; beliefs by Normal explanations.

Moser, P.K. 1990. Physicalism and intentional attitudes. *Behavior and Philosophy* 18:33-41.

Peacocke, C. 1983. Between instrumentalism and brain-writing. In *_Sense and Content_*. Oxford University Press.

Instrumentalism about belief can't be right, because of Martian marionettes, but the language of thought is too strong a requirement. A state's structured content may reside in its pattern of relations to other states.

Possin, K. 1986. The case against Stich's Syntactic Theory of Mind. *Philosophical Studies* 49:405-18.

Stich is wrong, circular, and representational anyway.

Pratt, I. 1993. Analysis and the attitudes. In (S. Wagner & R. Warner, eds) *Naturalism: A Critical Appraisal*. University of Notre Dame Press.

Pylyshyn, Z.W. 1987. What's in a mind? *Synthese* 70:97-122.

Must individuate mental states by semantics, not just by function.

Recanati, F. 1997. Can we believe what we do not understand? *Mind and Language* 12:84-100.

Robinson, W.S. 1990. States and beliefs. *Mind* 99:33-51.

Schwartz, J. 1992. Propositional attitude psychology as an ideal type. *Topoi* 11:5-26.

Smith, D.M. 1994. Toward a perspicuous characterization of intentional states. *Philosophical Studies* 74:103-20.

Sobel, D. & Copp, D. 2001. Against direction of fit accounts of belief and desire. *Analysis* 61:44-53.

Sperber, D. 1997. Intuitive and reflective beliefs. *Mind and Language* 12:67-83.

Stich, S.P. 1983. *From Folk Psychology to Cognitive Science*. MIT Press. Beliefs/desires are out, new Syntactic Theory is in.

Stich, S.P. 1984. Relativism, rationality, and the limits of intentional ascription. *Pacific Philosophical Quarterly*.

Stone, T. & Young, A.W. 1997. Delusions and brain injury: The philosophy and psychology of belief. *Mind and Language* 12:327-364.

Von Eckardt, B. & Poland, J. 2000. In defense of the standard view. *Protosociology* 14:312-331.

Weatherall, P. 1996. What do propositions measure in folk psychology? *Philosophical Psychology* 9:365-80.

2.1e The Nature of Folk Psychology

Blackburn, S. 1992. Theory, observation, and drama. *Mind and Language* 7:187.

Bogdan, R.G. (ed) 1991. *Mind and Common Sense: Philosophical Essays on Commonsense Psychology*. Cambridge University Press.

Botterill, G. 1996. Folk psychology and theoretical status. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Churchland, P.M. 1988. Folk psychology and the explanation of human behavior. *Proceedings of the Aristotelian Society* 62:209-21. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Folk psychology is a theory: defense against objections from logicality, softness of laws, practical function, behavior, and simulation. It needn't be a deductive-nomological theory; e.g. it might be based on prototypes.

Clark, A. 1987. From folk psychology to naive psychology. *Cognitive Science* 11:139-54.

Folk psychology isn't all that bad. It survived evolution after all.

Collins, J. 2000. Theory of mind, logical form and eliminativism. *Philosophical Psychology* 13:465-490.

Dennett, D.C. 1991. Two contrasts: Folk craft vs folk science and belief vs opinion. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

FP is craft, not theory. Opinions rather than beliefs are interesting.

Falvey, K. 1999. A natural history of belief. *Pacific Philosophical Quarterly* 80:324-345.

Fletcher, G. 1995. *The Scientific Credibility of Folk Psychology*. Lawrence Erlbaum.

Fletcher, G. 1995. Two uses of folk psychology: Implications for psychological science. *Philosophical Psychology* 8:375-88.

Goldman, A. 1992. The psychology of folk psychology. *Behavioral and Brain Sciences*.

On the psychology of self-ascription of mental states. Functionalism has serious problems, as we don't have direct access to causal roles. Defends a qualia-based account, even for propositional attitudes.

Gopnik, A. 1990. Developing the idea of intentionality: Children's theories of mind. *Canadian Journal of Philosophy* 20:89-114.

On the development of folk-psychological concepts in children. First the appearance/reality distinction, then more complex theories of perception, representation, and belief. Implications for the status of folk psychology.

Gopnik, A. & Wellman, H. 1992. Why the child's theory of mind really is a theory. *Mind and Language* 7:145-71.

Graham, G. 1987. The origins of folk psychology. *Inquiry* 30:357-79.

Greenwood, J.D. (ed) 1991. *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge University Press.

Leon, M. 1998. The unnaturalness of the mental: The status of folk psychology. *Southern Journal of Philosophy* 36:367-92.

Lycan, W.G. 1997. Folk psychology and its liabilities. In (M. Carrier & P. Machamer, eds) *Mindsapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Margolis, J. 1991. The autonomy of folk psychology. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

McDonough, R. 1991. A culturalist account of folk psychology. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

Morton, A. 1980. *Frames of Mind*. Oxford University Press.

Morton, A. 1991. The inevitability of folk psychology. In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.

Morton, A. 1996. Folk psychology is not a predictive device. *Mind* 105:119-37.

Pettit, P. 2000. How the folk understand folk psychology. *Protosociology* 14:26-38.

Place, U.T. 1996. Folk psychology from the standpoint of conceptual analysis. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Pratt, I. 1996. Encoding psychological knowledge. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Preston, J.M. 1989. Folk psychology as theory or practice? The case for eliminative materialism. *Inquiry* 32:277-303.

Defending the claim that folk psychology is an empirical pre-scientific theory, with its own laws. In a particular, a detailed reply to the criticisms in Wilkes 1984.

Robinson, W.S. 1996. Mild realism, causation, and folk psychology. *Philosophical Psychology* 8:167-87.

Schwitzgebel, E. 2001. In-between believing. *Philosophical Quarterly* 51:76 - 82.

Sehon S.R. 1997. Natural kind terms and the status of folk psychology. *American Philosophical Quarterly* 34:333-44.

Sharpe, R. 1987. The very idea of a folk psychology. *Inquiry* 30:381-93.

Smith, B.C. 1996. Does science underwrite our folk psychology? In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Stemmer, N. 1995. A behaviorist account to theory and simulatytion theories of folk psychology. *Behavior and Philosophy* 23:29-41.

Sterelny, K. 1998. Intentional agency and the metarepresentation hypothesis. *Mind and Language* 13:11-28.

Stich, S.P. & Ravenscroft, R. 1994. What is folk psychology? *Cognition* 50:447-68. Reprinted in (Stich) *Deconstructing the Mind*. Oxford University Press, 1996.

Distinguishes internal and external accounts of folk psychology (mechanisms vs systematizations), and various versions of each of these. Only some are compatible with eliminativist arguments.

von Eckardt, B. 1997. The empirical naivete in the current philosophical conception of folk psychology. In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Wilkes, K.V. 1984. Pragmatics in science and theory in common sense. *Inquiry* 27:339-61.

Wilkes, K.V. 1991. The relationship between scientific psychology and common-sense psychology. *Synthese* 89:15-39.

Common-sense psychology is no theory at all, and not in competition with scientific psychology. CSP is particular, rich, vague; SP is general, austere, precise. CSP will be neither subsumed nor eliminated by SP.

Wilkes, K.V. 1991. The long past and the short history. In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.

Argues that commonsense and scientific psychology are quite distinct in their aims, scope, framework, and nature, but have been confused by philosophy. With support from historical considerations.

2.1f The Simulation Theory

Arkway, A. 2000. The simulation theory, the theory theory and folk psychological explanation. *Philosophical Studies* 98:115-137.

Carruthers, P. 1996. Simulation and self-knowledge: A defence of the theory-theory. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Carruthers, P. & Smith, P. 1996. *Theories of Theories of Mind*. Cambridge University Press.

Cruz, J.L.H. 1998. Mindreading: Mental state ascription and cognitive architecture. *Mind and Language* 13:323-340.

Currie, G. 1995. Visual imagery as the simulation of vision. *Mind and Language* 10:25-44.

Currie, G. 1996. Simulation-theory, theory-theory, and the evidence from autism. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Currie, G. & Ravenscroft, I. 1997. Mental simulation and motor imagery. *Philosophy of Science* 64:161-80/

Currie, G. 1998. Pretence, pretending, and metarepresenting. *Mind and Language* 13:35-55.

Davies, M. 1992. The mental simulation debate. In (E. Villanueva, ed) *Truth and Rationality*. Ridgeview.

Davies, M. & Stone, T. (eds) 1995. *Folk Psychology: The Theory of Mind Debate*. Blackwell.

Davies, M. & Stone, T. (eds) 1995. *Mental Simulation: Evaluations and Applications*. Blackwell.

Freeman, N.H. 1995. Theories of mind in collision: Plausibility and authority. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Fuller, G. 1995. Simulation and psychological concepts. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Goldman, A. 1989. Interpretation psychologized. *Mind and Language* 4:161-85. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Goldman, A. 1992. In defense of the simulation theory. *Mind and Language* 7:104-119. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Goldman, A. 1996. Simulation and interpersonal utility. In (L. May, M. Friedman, & A. Clark, eds) *Mind and Morals: Essays on Ethics and Cognitive Science*. MIT Press.

Goldman, A. 2000. Folk psychology and mental concepts. *Protosociology* 14:4-25.

Gopnik, A. & Wellman, H.M. 1995. Why the child's theory of mind really is a theory. *Mind and Language*. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Gopnik, A. and Meltzoff, AN. 1998. Theories vs. modules: To the max and beyond. A reply to Poulin-Dubois and to Stich and Nichols. *Mind and Language* 13:450-456.

Gordon, R.M. 1986. Folk psychology as simulation. *Mind and Language* 1:158-71.

Reprinted in (M. Davies & T. Stone, eds) *_Folk Psychology_*. Blackwell.

FP is a strategy for prediction via simulation; an ability, not a theory.

Gordon, R.M. 1992. The simulation theory: objections and misconceptions. *Mind and Language* 7:11-34. Reprinted in (M. Davies & T. Stone, eds) *_Folk Psychology_*. Blackwell.

Gordon, R.M. & Barker, J.A. 1994. Autism and the "theory of mind" debate. In (G. Graham & G.L. Stephens, eds) *_Philosophical Psychopathology_*. MIT Press.

Gordon, R.M. 1995. Simulation without introspection or inference from me to you. In (M. Davies & T. Stone, eds) *_Mental Simulation_*. Blackwell.

Gordon, R.M. 1996. Sympathy, simulation, and the impartial spectator. In (L. May, M. Friedman, & A. Clark, eds) *_Mind and Morals: Essays on Ethics and Cognitive Science_*. MIT Press.

Gordon, R.M. 1996. 'Radical' simulationism. In (P. Carruthers & P. Smith, eds) *_Theories of Theories of Mind_*. Cambridge University Press.

Gordon, R.M. 2000. Sellars's Rylean ancestors revisited. *Protosociology* 14:102-114.

Greenwood, J.D. 1999. Simulation, theory-theory and cognitive penetration: No "instance of the fingerpost". *Mind and Language* 14:32-56.

Heal, J. 1986. Replication and functionalism. In (J. Butterfield, ed) *_Language, Mind, and Logic_*. Cambridge University Press. Reprinted in (M. Davies & T. Stone, eds) *_Folk Psychology_*. Blackwell.

Heal, J. 1994. Simulation vs. theory-theory: What is at issue? In (C. Peacocke, ed) *_Objectivity, Simulation, and the Unity of Consciousness_*. Oxford University Press.

Heal, J. 1995. How to think about thinking. In (M. Davies & T. Stone, eds) *_Mental Simulation_*. Blackwell.

Heal, J. 1996. Simulation and cognitive penetrability. *Mind and Language* 11:44-67.

Heal, J. 1996. Simulation, theory, and content. In (P. Carruthers & P. Smith, eds) *_Theories of Theories of Mind_*. Cambridge University Press.

Heal, J. 1998. Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind and Language* 13:477-498.

Heal, J. 2000. Understanding other minds from the inside. *Protosociology* 14:39-55.

Henderson, D. 1996. Simulation theory versus theory theory: A difference without a difference in explanations. *Southern Journal of Philosophy* 34:65-93.

Kuhberger, A, Perner, J., Schulte, M., & Leingruber, R. 1995. Choice or no choice: Is the Langer effect evidence against simulation? *Mind and Language* 10:423-36.

Leslie, A.M. & German, T.P. 1995. Knowledge and ability in "theory of mind": A one-eyed overview of a debate. In (M. Davies & T. Stone, eds) *_Mental Simulation_*. Blackwell.

Levin, J. 1995. Folk psychology and the simulationist challenge. *Acta Analytica* 10:77-100.

Nichols, S., Stich, S., & Leslie, A. 1995. Choice effects and the

ineffectiveness of simulation. *Mind and Language* 10:437-45.

Nichols, S., Stich, S., Leslie, A., Klein, D. 1996. Varieties of off-line simulation. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Nichols, S. & Stich, S. 1998. Rethinking co-cognition: A reply to Heal. *Mind and Language* 13:499-512.

Perner, J. 1994. The necessity and impossibility of simulation. In (C. Peacocke, ed) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford University Press.

Perner, J. 1996. Simulation as explicitation of predication-implicit knowledge about the mind: Arguments for a simulation-theory mix. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Perner, J., Gschaidner, A., Kuhberger, A. & Schrofner, S. 1999. Predicting others through simulation or by theory? A method to decide. *Mind and Language* 14:57-79.

Pust, J. 1999. External accounts of folk psychology, eliminativism, and the simulation theory. *Mind and Language* 14:113-130.

Ruffman, T. 1996. Do children understand the mind by means of a simulation or a theory? Evidence from their understanding of inference. *Mind and Language* 11:388-414.

Scholl, B.J. & Leslie, A.M. 1999. Modularity, development and "theory of mind." *Mind and Language* 14:131-153.

Schwitzgebel, E. 1999. Representation and desire: a philosophical error with consequences for theory-of-mind research. *Philosophical Psychology* 12:157-180.

Stich, S.P. & Nichols, S. 1993. Folk psychology: simulation or tacit theory? *Mind and Language* 7:35-71. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Stich, S.P. & Nichols, S. 1995. Second thoughts on simulation. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Stich, S.P. & Nichols, S. 1997. Cognitive penetrability, rationality, and restricted simulation. *Mind and Language* 12:297-326.

Stich, S. & Nichols, S. 1998. Theory theory to the max. *Mind and Language* 13:421-449.

Stone, T. & Davies, M. 1996. The mental simulation debate: A progress report. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Wilkerson, W.S. 2001. Simulation, theory, and the frame problem: the interpretive moment. *Philosophical Psychology* 14:141-153.

2.2 Internalism and Externalism [see also 1.5d]

2.2a Is Content in the Head? (Putnam, Burge)

Antony, M. 1993. Social relations and the individuation of thought. *Mind* 102:247-61.

Bilgrami, A. 1987. An externalist account of psychological content.
Philosophical Topics 15:191-226.

Developing an externalist account consistent with psychological explanation.
Contra Burge, social links aren't constitutive of content. Causal links
are indirectly constitutive of content, via our conceptions.

Brueckner, A. 1995. The characteristic thesis of anti-individualism. Analysis
55:146-48.

Bruns, M. & Soldati, G. 1997. Object-dependent and property-dependent
concepts. Dialectica 48:185-208.

Burge, T. 1979. Individualism and the mental. Midwest Studies in Philosophy
4:73-122.

Belief contents are not fully determined by internal state, as the linguistic
community plays an important role: arthritis, brisket, contract, sofa, etc.
So mental states are not individuated individualistically.

Burge, T. 1982. Other bodies. In (A. Woodfield, ed) _Thought and Object_.
Oxford University Press.

On Putnam's Twin Earth. Natural kind terms are not indexical. Even de dicto
attitudes are not in the head; they presuppose the existence of other things.

Burge, T. 1986. Intellectual norms and foundations of mind. Journal of
Philosophy 83:697-720.

On non-individualist elements due to by intellectual norms in the community,
to which meanings are answerable. Even meaning-giving truths can be doubted.
With remarks on sofas/safos, and on linguistic meaning vs. cognitive value.

Butler, K. 1993. Individualism, computationalism, and folk psychology.
Manuscript.

Challenges Burge's interpretations of the thought-experiments: e.g. twins
have the same concept, neither of which is the public concept of arthritis.
With remarks on computationalism and Marr's theory.

Campbell, J. 1982. Extension and psychic state: Twin Earth revisited.
Philosophical Studies 42:67-89.

Argues that natural kind terms are token-reflexive, with reference ultimately
fixed to the underlying explanatory properties of the surface qualities of
local matter.

Crane, T. 1991. All the difference in the world. Philosophical Quarterly
41:1-25.

Twins share the same concepts. Contra Putnam: essentialism is fallacious;
contra Burge: speakers share beliefs, but one has false belief about meaning.

Cummins, R. 1991. Methodological reflections on belief. In (R. Bogdan, ed)
Mind and Common Sense. Cambridge University Press.

We shouldn't rely on intuitions about thought-experiments; we need an
empirical theory about belief. Belief contents are distinct from sentence
contents; we have to distinguish linguistic from psychological semantics.

Devitt, M. 1990. Meanings just ain't in the head. In (G. Boolos, ed) _Meaning
and Method: Essays in Honor of Hilary Putnam_. Cambridge University Press.

Against Searle's theory of internal intentionality. Searle's theory
requires magic to grasp external contents internally.

Dretske, F. 1993. The nature of thought. Philosophical Studies 70:185-99.

Argues that thought is extrinsic, but it is not essentially social. A
system without a linguistic community could have thoughts, if it had an
appropriate learning history.

Elugardo, R. 1993. Burge on content. Philosophy and Phenomenological Research

53:367-84.

Contra Burge on sofas: oblique that-clauses can't identify the (wide) way that the subject thinks of sofas, which is idiosyncratic and inexpressible.

Forbes, G. 1987. A dichotomy sustained. *Philosophical Studies* 51:187-211.

Gives a Fregean account of belief semantics to handle the Burge cases, and argues that the *type* of a proposition may be internal even if the token itself is not. With remarks on the relevance to Grice's program.

Georgalis, N. 1999. Rethinking Burge's thought experiment. *Synthese* 118:145-64.

Horowitz, A. 1995. Putnam, Searle, and externalism. *Philosophical Studies* 81:27-69.

Argues for a moderate externalism by synthesizing Putnam and Searle: internal intension leaves extension indeterminate, but it specifies the facts relevant to filling in the indeterminacy.

Koethe, J. 1992. And they ain't outside the head either. *Synthese* 90:27-53.

Ludwig, K. 1993. Externalism, naturalism, and method. In (E. Villanueva, ed) *Naturalism and Normativity*. Ridgeview.

Ludwig, K. 1996. Duplicating thoughts. *Mind and Language* 11:92-102.

Mandelkar, S. 1991. An argument against the externalist account of psychological content. *Philosophical Psychology* 4:375-82.

Argues that conscious experience is required for intentional states, and that any external relations could be satisfied without this experience, so external relations cannot suffice for intentional content.

McCulloch, G. 1992. The spirit of twin earth. *Analysis* 52:168-174.

Various arguments against Crane 1991 on externalism.

McDowell, J. 1977. On the sense and reference of a proper name. *Mind*.

McGilvray, J. 1998. Meanings are syntactically individuated and found in the head. *Mind and Language* 13:225-280.

McKinsey, M. 1991. The internal basis of meaning. *Pacific Philosophical Quarterly* 72:143-69.

Argues that meaning is determined by a certain kind of internal state, involving de se cognitive attitudes. These states aren't shared by twins, but are still narrow in a strong sense.

McKinsey, M. 1993. Curing folk psychology of arthritis. *Philosophical Studies* 70:323-36.

McKinsey, M. 1994. Individuating beliefs. *Philosophical Perspectives* 8:303-30.

Owens, J. 1983. Functionalism and the propositional attitudes. *Nous* 17:529-49.

Functional organization doesn't determine attitude content, even if we include inputs and outputs.

Perry, J. 1979. The problem of the essential indexical. *Nous* 13:3-21.

Indexicals are essential to some beliefs, so belief cannot just be a relation to a proposition. Belief contents must be at least in part construed relative to a subject. Separate belief object and belief state.

Putnam, H. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7:131-193. Reprinted in *Mind, Language, and Reality*.

(Cambridge University Press, 1975).

What is in the head doesn't determine the reference of our thoughts: my twin on Twin Earth refers to XYZ where I refer to H₂O. Content is determined by environment and linguistic community as well as by internal stereotypes.

Putnam, H. 1987. Meaning, other people, and the world. In *Representation and Reality*. MIT Press.

Meanings **still** aren't in the head.

Searle, J.R. 1983. *Intentionality*. Cambridge University Press.

Sure, meanings **are** in the head -- e.g. the content of a given visual experience is "the thing that is causing this experience".

Sosa, E. 1991. Between internalism and externalism. In (E. Villanueva, ed) *Consciousness*. Ridgeview.

Sosa, E. 1993. Abilities, concepts, and externalism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

On concepts as abilities, and on construals of abilities that lead to internalism and externalism. Maybe the relevant abilities are characterized externally but determined internally. Remarks on Putnam, Davidson, Burge.

Stalnaker, R. 1993. Twin earth revisited. *Proceedings of the Aristotelian Society* 63:297-311.

Making sense of twin earth intuitions with an information-theoretic account of content: information depends on relations in normal conditions, which are extrinsic. With remarks on the context-sensitivity of content-attribution.

Wikforss, A. 2001. Social externalism and conceptual errors. *Philosophical Quarterly* 203:217-31.

Woodfield, A. 1982. Thought and the social community. *Inquiry* 25:435-50.

Burge's arguments show only that context-ascription is pragmatically sensitive to context, depending on the epistemic predicament of the ascriber. Content itself is still internal.

Zemach, E.M. 1976. Putnam's theory on the reference of substance terms. *Journal of Philosophy* 73:116-27.

Argues that the extension of 'water' is the same on earth and twin earth, using arguments from isotopes and scientific development. Molar properties determine classification. Remarks on historicism and the division of labor.

2.2b Externalism and Psychological Explanation (Burge, Fodor)

Adams, F. & K. Aizawa, 2001. The bounds of cognition. *Philosophical Psychology* 14:43-64.

Arjo, D. 1996. Sticking up for Oedipus: Fodor on intentional generalizations and broad content. *Mind and Language* 11:231-45.

Buller, D.J. 1992. "Narrow"-minded breeds inaction. *Behavior and Philosophy* 20:59-70.

Buller, D.J. 1997. Individualism and evolutionary psychology (or: In defense of "narrow" functions). *Philosophy of Science* 64:74-95.

Burge, T. 1982. Two thought experiments reviewed. *Notre Dame Journal of Formal Logic* 23:284-94.

Reply to Fodor 1982, clarification of position.

Burge, T. 1986. Individualism and psychology. *Philosophical Review* 95:3-45.

Psychology should be and is done non-individualistically, i.e. with reference

to environment. Examples from vision, e.g. Marr.

Clark, A. & Chalmers, D.J. 1998. The extended mind. *Analysis* 58:7-19.
Advocates a different sort of "active externalism", based on the role of the environment in actively driving cognition. Beliefs can extend into an agent's immediate environment (e.g. a notebook) in this way.

Dretske, F. 1992. What isn't wrong with folk psychology. *Metaphilosophy* 23:1-13.
Argues that extrinsic properties can play a respectable role in scientific explanation; e.g. the histories of plants, animals, and devices are relevant in explaining their current behavior.

Egan, F. 1991. Must psychology be individualistic? *Philosophical Review* 100:179-203.
Maybe, maybe not. Contra Fodor: science can be non-individualistic.
Contra Burge re oblique ascriptions and Marr.

Fodor, J.A. 1980. Methodological solipsism as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3:63-109. Reprinted in *_RePresentations_* (MIT Press, 1980).
Should do psychology without reference to the external world. What counts for psychology is in the head; who cares about truth, reference, and the rest?

Fodor, J.A. 1982. Cognitive science and the twin-earth problem. *Notre Dame Journal of Formal Logic* 23:98-118.
Twin Earth isn't a problem for cognitive science. Intents of utterances, de re/de dicto, etc. Truth conditions aren't in the head, but that's no problem.

Gauker, C. 1987. Mind and chance. *Canadian Journal of Philosophy* 17:533-52.

Globus, G. 1984. Can methodological solipsism be confined to psychology? *Cognition and Brain Theory* 7:233-46.
Methodological solipsism implies epistemological solipsism.

Hardcastle, V.G. 1997. [Explanation] is explanation better. *Philosophy of Science* 64:154-60.

Hurley, S.L. 1998. Vehicles, contents, conceptual structure, and externalism. *Analysis* 58:1-6.

Jacob, P. 1993. Externalism and the explanatory relevance of broad content. *Mind and Language* 8:131.

Kitcher, P.S. 1984. Narrow taxonomy and wide functionalism. *Philosophy of Science* 52:78-97.
Argues against Stich, Fodor, Block: use different taxonomies (narrow/wide) for different purposes. Both are OK, functionalism *and* content survive.

Kobes, B. 1989. Semantics and psychological prototypes. *Pacific Philosophical Quarterly* 70:1-18.
Relates the individualism debate to Roschian prototype research.

Losonsky, M. 1995. Emdeded systems vs. individualism. *Minds and Machines* 5:357-71.

Macdonald, C. 1992. Weak externalism and psychological reduction. In (D Charles & K. Lennon, eds) *_Reduction, Explanation and Realism_*. Oxford University Press.

Marras, A. 1985. The Churchlands on methodological solipsism and computational psychology. *Philosophy of Science* 52:295-309.
MS doesn't rule out all use of content, just of wide content. Narrow content

is OK. With remarks on folk psychology and computation.

Maloney, J.C. 1985. Methodological solipsism reconsidered as a research strategy in cognitive psychology. *Philosophy of Science* 52:451-69.

Various problems for computational psychology handling content. It shares the problems of a naturalistic psychology.

McClamrock, R. 1991. Methodological individualism considered as a constitutive principle of scientific inquiry. *Philosophical Psychology* 4:343-54.

McClamrock, R. 1995. *_Existential Cognition: Computational Minds in the World_*. University of Chicago Press.

Noonan, H.W. 1984. Methodological solipsism: A reply to Morris. *Philosophical Studies* 48:285-290.

Noonan, H.W. 1986. Russellian thoughts and methodological solipsism. In (J. Butterfield, ed) *_Language, Mind, and Logic_*. Cambridge University Press.

Noonan, H.W. 1990. Object-dependent thoughts and psychological redundancy. *Analysis* 51:1-9.

Noonan, H.W. 1993. Object-dependent thoughts: A case of superficial necessity but deep contingency? In (J. Heil & A. Mele, eds) *_Mental Causation_*. Oxford University Press.

Object-dependent thoughts are redundant in psychological explanation, as an explanation applying to a hallucinator will work as well. But this needn't defeat externalism in general. With remarks on self-knowledge.

Patterson, S. 1990. The explanatory role of belief ascriptions. *Philosophical Studies* 59:313-32.

Uses examples to argue that in explaining behavior we often ascribe beliefs in an individualistic way, even in cases where individual and community use diverge. These contents are at least sometimes expressible.

Patterson, S. 1991. Individualism and semantic development. *Philosophy of Science* 58:15-35.

Developmental psychologists attribute concepts individualistically.

Peacocke, C. 1993. Externalist explanation. *Proceedings of the Aristotelian Society* 67:203-30.

Externalist states are required for the explanation of relational properties. Counters objections from conceptual connections and dormitive-virtue worries, and applies to teleology, self-knowledge, etc.

Petrie, B. 1990. Nonautonomous psychology. *Southern Journal of Philosophy* 28:539-59.

Argues that behavior is often individuated widely for explanatory purposes, so that wide content is relevant, and that there is more to causation than local causation, so Stich's autonomy principle fails.

Pettit, P. 1986. Broad-minded explanation and psychology. In (P. Pettit & J. McDowell, eds) *_Subject, Thought and Context_*. Oxford University Press.

Rowlands, M. 1995. Against methodological solipsism: The ecological Approach. *Philosophical Psychology* 8:5-24.

Segal, G. 1989. The return of the individual. *Mind* 98:39-57.

Sterelny, K. 1990. Animals and individualism. In (P. Hanson, ed) *_Information, Language and Cognition_*. University of British Columbia Press.

Stich, S.P. 1978. Autonomous psychology and the belief/desire thesis. *Monist*

61:573-91. Reprinted in (W. Lycan, ed) Mind and Cognition (Blackwell, 1990).
Beliefs are not in the head, so aren't good for psychological explanation.
Interesting, but confuses the role of truth-values with truth-conditions.

Tuomela, R. 1989. Methodological solipsism and explanation in psychology.
Philosophy of Science 56:23-47.

Wallace, J. & Mason, H.E. 1990. On some thought experiments about mind and
meaning. In (C. Anderson & J. Owens, eds) Propositional Attitudes. CSLI.

Wilson, R.A. 1994. Causal depth, theoretical appropriateness, and
individualism in psychology. *Philosophy of Science* 61:55-75.

Wilson, R.A. 1995. Cartesian Psychology and Physical Minds: Individualism and
the Sciences of the Mind. Cambridge University Press.

2.2c Externalism and Mental Causation

Adams, F. 1993. Fodor's modal argument. *Philosophical Psychology* 6:41-56.

Allen, C. 1995. It isn't what you think: A new idea about intentional
causation. *Nous* 29:115-26.

Baker, L.R. 1994. Content and context. *Philosophical Perspectives* 8:17-32.
Argues contra Fodor that broad contents can be explanatory -- if they can't,
no relational properties can. Fodor's "no-conceptual-connection" and
"cross-context" tests for causal powers fail to do the job.

Barrett, J. 1997. Individualism and the cross-contexts test. *Pacific
Philosophical Quarterly* 78:242-??.

Braun, D. 1991. Content, causation, and cognitive science. *Australasian
Journal of Philosophy* 69:375-89.
Arguments for the causal significance of broad content. Physical twins can
differ in causal powers; broad content figures in (ceteris paribus) causal
generalizations; can make twin arguments against narrow content too. Hmm.

Burge, T. 1989. Individuation and causation in psychology. *Pacific
Philosophical Quarterly* 70:303-22.
Contra Fodor: psychological processes can play differing causal roles,
despite being physically identical.

Burge, T. 1995. Intentional properties and causation. In (C. Macdonald & G.
Macdonald, eds) Philosophy of Psychology: Debates about Psychological
Explanation. Blackwell.
Reply to Fodor 1991.

Butler, K. 1996. Content, causal powers, and context. *Philosophy of Science*
63:105-14.

Christensen, D. 1992. Causal powers and conceptual connections. *Analysis*
52:163-8.
Fodor's modal argument for narrow content rests on a false analogy between
cases concerning thoughts and those concerning planets.

Fodor, J.A. 1991. A modal argument for narrow content. *Journal of Philosophy*
88:5-26.
On when a difference in effects amounts to a difference in causal powers:
when the effects are connected contingently, not conceptually, to the causes.
Differences in wide content don't satisfy this, so aren't causal powers.

Garcia-Carpintero, M. 1994. The supervenience of mental content. *Proceedings*

of the Aristotelian Society 68:117-135.

Mental content can be extrinsic and efficacious. Narrow content strategies don't work, as observation concepts are still extrinsic. One can't screen of the intrinsic part from the rest. Thought-experiments are inconclusive.

Heil, J. & Mele, A. 1991. Mental causes. *American Philosophical Quarterly* 28:61-71.

Reconciling Twin Earth with the causal relevance of content. Historical factors can be causally relevant.

Jacob, P. 1992. Externalism and mental causation. *Proceedings of the Aristotelian Society* 66:203-19.

Argues that externalist content is not causally efficacious, but is relevant to causal explanations of behavior indirectly, via the cognitive activities of others external to the system.

Klein, M. 1996. Externalism, content, and causation. *Proceedings of the Aristotelian Society* 96:159-76.

Lalor, B.J. 1997. It is what you think: intentional potency and anti-individualism. *Philosophical Psychology* 10:165-78.

Ludwig, K. 1993. Causal relevance and thought content. *Philosophical Quarterly* 44:334-53.

McGinn, C. 1991. Conceptual causation. *Mind* 100:525-46.

Montgomery, R. 1995. Non-Cartesian explanations meet the problem of mental causation. *Southern Journal of Philosophy* 33:221-41.

Owens, J. 1993. Content, causation, and psychophysical supervenience. *Philosophy of Science* 60:242-61.

Russow, L.M. 1993. Fodor, Adams, and causal properties. *Philosophical Psychology* 6:57-61.

Saidel, E. 1994. Content and causal powers. *Philosophy of Science* 61:658-65.

Segal, G. & Sober, E. 1991. The causal efficacy of content. *Philosophical Studies* 63:1-30.

Seymour, D. 1993. Some of the difference in the world: Crane on intentional causation. *Philosophical Quarterly* 43:83-89.

Sturgeon, S. 1994. Good reasoning and cognitive architecture. *Mind and Language* 9:88-101.

Epistemology requires the causal relevance of content, and the relevant content is narrow. On how various architectures might support this causal relevance, by being realized by more specific intrinsic features.

van Gulick, R. 1989. Metaphysical arguments for internalism and why they don't work. In (S. Silvers, ed) *_ReRepresentation_*. Kluwer.

Against some arguments for internalism: local causation doesn't imply local type-individuation, as distal relations affect distal causes and effects; and processes can have access to semantic properties via formal properties.

Wilson, R.A. 1992. Individualism, causal powers, and explanation. *Philosophical Studies* 68:103-39.

Science frequently appeals to relational and historical taxonomies, so either causal powers can be non-intrinsic or science needn't taxonomize by causal powers. With remarks on causal properties and conceptual connections.

Wilson, R.A. 1993. Against a priori arguments for individualism. *Pacific*

Philosophical Quarterly 74:60-79.

Arguments from causal powers beg the question, either on whether relational properties can have causal powers or on whether science taxonomizes by causal powers, as relational properties are common in scientific explanation.

Yablo, S. 1997. Wide causation. Philosophical Perspectives 11:251-81.

2.2d Externalism and the Theory of Vision

Burge, T. 1986. Individualism and psychology. Philosophical Review 95:3-45.

Psychology should be and is done non-individualistically, i.e. with reference to environment. Examples from vision, e.g. Marr.

Butler, K. 1996. Individualism and Marr's computational theory of vision. Mind and Language 11:313-37.

Butler, K. 1996. Content, computation, and individualism in vision theory. Analysis 56:146-54.

Cain, M.J. 2000. Individualism, twin scenarios and visual content. Philosophical Psychology 13:441-463.

Davies, M. 1991. Individualism and perceptual content. Mind 100:461-84.

Egan, F. 1992. Individualism, computation, and perceptual content. Mind 101:443-59.

Egan, F. 1996. Intentionality and the theory of vision. In (K. Akins, ed) *_Perception_*. Oxford University Press.

Francescotti, R.M. 1991. Externalism and the Marr theory of vision. British Journal for the Philosophy of Science 42:227-38.

Kitcher, P.S. 1988. Marr's computational theory of vision. Philosophy of Science 55:1-24.

Morton, P. 1993. Supervenience and computational explanation in vision theory. Philosophy of Science 60:86-99.

Patterson, S. 1996. Success-orientation and individualism in the theory of vision. In (K. Akins, ed) *_Perception_*. Oxford University Press.

Segal, G. 1989. Seeing what is not there. Philosophical Review 97:189-214.
Contra Burge, Marr's theory is individualistic. Intentional contents therein are neutral between twins' environments; nothing grounds a more specific attribution.

Segal, G. 1991. Defence of a reasonable individualism. Mind 100:485-94.

Shapiro, L.A. 1993. Content, kinds, and individualism in Marr's theory of vision. Philosophical Review 102:489-513.

Contra Segal, Marr's theory is non-individualistic even though it may classify twins together. Computational-level task descriptions rather than behavior guide content ascription, so the environment plays a crucial role.

Shapiro, L.A. 1997. A clearer vision. Philosophy of Science 64:131-53.

Shapiro, L.A. 1997. Junk representations. British Journal for the Philosophy of Science .

2.2e Externalism and Computation

Andler, D. 1995. Can we knock off the shackles of syntax? In (E. Villanueva, ed) *_Content_*. Ridgeview.

Butler, K. 1998. Content, computation, and individuation. *Synthese* 114:277-92.

Egan, F. 1995. Computation and content. *Philosophical Review* 104:181-203.

Egan, F. 1999. In defence of narrow mindedness. *Mind and Language* 14:177-94.

Kazez, J.R. 1994. Computationalism and the causal role of content. *Philosophical Studies* 75:231-60.

Kobes, B. 1990. Individualism and artificial intelligence. *Philosophical Perspectives* 4:429-56.

Winograd's SHRDLU doesn't support individualism: its concepts are anchored (to a fictional world) via its programmer, and it could have made errors.

Miscevic, N. 1996. Computation, content, and cause. *Philosophical Studies* 82:241-63.

Peacocke, C. 1995. Content, computation, and externalism. In (E. Villanueva, ed) *_Content_*. Ridgeview.

Peacocke, C. 1999. Computation as involving content: A response to Egan. *Mind and Language* 14:195-202.

Seager, W.E. 1992. Thought and syntax. *Philosophy of Science Association* 1992, 1:481-91.

Syntax is extrinsically determined, as well as semantics. So if broad content is irrelevant to psychology, syntax is too.

Wilson, R.A. 1994. Wide computationalism. *Mind* 103:351-72.

2.2f Externalism and Self-Knowledge

Berg, J. 1998. First-person authority, externalism, and wh-knowledge. *Dialectica* 52:41-44.

Bernecker, S. 1996. Davidson on first-person authority and externalism. *Inquiry* 39:121-39.

Bernecker, S. 1996. Externalism and the attitudinal component of self-knowledge. *Nous* 30:262-75.

Bernecker, S. 1998. Self-knowledge and closure. In (P. Ludlow & N. Martin, eds) *_Externalism and Self-Knowledge_*. CSLI.

Bilgrami, A. 1992. Can externalism be reconciled with self-knowledge? *Philosophical Topics* 20:233-68.

Boghossian, P. 1989. Content and self-knowledge. *Philosophical Topics* 17:5-26.

We can't know our thought-contents by inference (circular), by introspection (because they're relational), or directly, so we can't know them at all.

Boghossian, P. 1992. Externalism and inference. *Philosophical Issues* 2:11-28.

Boghossian, P. 1994. The transparency of mental content. *Philosophical Perspectives* 8:33-50.

Boghossian, P. 1997. What the externalist can know a priori. *Proceedings of*

the Aristotelian Society 97:161-75.

Brown, J. 1995. The incompatibility of anti-individualism and privileged access. *Analysis* 55:149-56.

Brown, J. 2000. Critical reasoning, understanding and self-knowledge. *Philosophy and Phenomenological Research* 61:659-676.

Brueckner, A. 1990. Scepticism about knowledge of content. *Mind* 99:447-51.

Brueckner, A. 1992. What an anti-individualist knows a priori. *Analysis* 52:111-18.

Contra McKinsey 1991, anti-individualism doesn't lead to a priori knowledge. The belief that water is wet doesn't conceptually entail facts about the external world (e.g. H₂O), although it may metaphysically necessitate them.

Brueckner, A. 1992. Semantic answers to skepticism. *Pacific Philosophical Quarterly* 73:200-19.

Brueckner, A. 1993. Skepticism and externalism. *Philosophia* 22:169-71.

Brueckner, A. 1994. Knowledge of content and knowledge of the world. *Philosophical Review*:103-327-43.

Brueckner, A. 1995. Trying to get outside your own skin. *Philosophical Topics* 23:79-111.

Brueckner, A. 1997. Externalism and memory. *Pacific Philosophical Quarterly* 78:1-12.

Brueckner, A. 1997. Is scepticism about self-knowledge incoherent? *Analysis* 4:287-90.

Brueckner, A. 2000. Externalism and the a prioricity of self-knowledge. *Analysis* 60:132-136.

Burge, T. 1988. Individualism and self-knowledge. *Journal of Philosophy* 85:649-63.

Knowledge of our thoughts is compatible with externalism: its content is self-referential and self-verifying. We needn't be able to explicate the content or its enabling conditions, or rule out twin possibilities.

Burge, T. 1996. Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society* 96:91-116.

Burge, T. 1998. Memory and self-knowledge. In (P. Ludlow & N. Martin, eds) *_Externalism and Self-Knowledge_*. CSLI.

Butler, K. 1997. Externalism, internalism, and knowledge of content. *philosophy and Phenomenological Research* 57:773-800.

Butler, K. 1998. Externalism and skepticism. *Dialogue* 37:13-34.

Chase, J. 2001. Is externalism about content inconsistent with internalism about justification? *Australasian Jouenal of Philosophy* 79:227-46.

Davidson, D. 1987. Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*.

Davies, M. 1998. Externalism, architecturalism, and epistemic warrant. In (C. Wright, B. Smith, and C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.

Ebbs, G. 1996. Can we take our words at face value? *Philosophy and*

Phenomenological Research 56:499-530.

Edwards, J. 1998. The simple theory of colour and the transparency of sense experience. In (C. Wright, B. Smith, and C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Falvey, K. & Owens, J. 1994. Externalism, self-knowledge, and skepticism. *Philosophical Review* 103:107-37.

Falvey, K. 2000. The compatibility of anti-individualism and privileged access. *Analysis* 60:137-142.

Gallois, A. 1994. Deflationary self-knowledge. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind: The Place of Philosophy in the Study of Mind*. Kluwer.

Gallois, A. & O'Leary-Hawthorne, J. 1996. Externalism and skepticism. *Philosophical Studies* 81:1-26.

Externalist anti-skeptical arguments fail as they require us to know a priori that our terms designate natural kinds, and also because they require us to know a priori that externalism is true. A thorough analysis.

Georgalis, N. 1990. No access for the externalist: Discussion of Heil's "Privileged access". *Mind* 100:101-8.

Georgalis, N. 1994. Asymmetry of access to intentional states. *Erkenntnis* 40:185-211.

Gibbons, J. 1996. Externalism and knowledge of content. *Philosophical Review* 105:287-310.

Gibbons, J. 2001. Externalism and knowledge of the attitudes. *Philosophical Quarterly* 51:13-28.

Glock, H.J. & Preston, J.M. 1995. Externalism and first-person authority. *Monist* 78:515-33.

Goldberg, S. 1997. Self-ascription, self-knowledge, and the memory argument. *Analysis* 57:211-19.

Goldberg, S. 1999. The relevance of discriminatory knowledge of content. *Pacific Philosophical Quarterly* 80:136-56.

Goldberg, S. 1999. The psychology and epistemology of self-knowledge. *Synthese* 118:165-201.

Goldberg, S. 2000. Externalism and authoritative knowledge of content: A new incompatibilist strategy. *Philosophical Studies* 100:51-79.

Hall, L. 1998. The self-knowledge that externalists leave out. *Southwest Philosophy Review* 14.

Heal, J. 1998. Externalism and memory. *Aristotelian Society Supplement* 72:77-94.

Heil, J. 1988. Privileged access. *Mind* 98:238-51.

Kobes, B. 1996. Mental content and hot self-knowledge. *Philosophical Topics* 24:71-99.

LePore, E. 1990. Subjectivism and environmentalism. *Inquiry* 33:197-214.
Subjectivism and environmentalism seem to clash on knowledge of content, but it's OK: under environmentalism we still know our contents w/o evidence.

- Ludlow, P. 1995. Externalism, self-knowledge, and the prevalence of slow-switching. *Analysis* 55:45-49.
Argues that cases of switching between language communities are quite common, so that Warfield's case for externalist self-knowledge doesn't work.
- Ludlow, P. 1995. Social externalism, self-knowledge, and memory. *Analysis* 55:157-59.
- Ludlow, P. 1995. Social externalism and memory: A problem? *Acta Analytica* 10:69-76.
- Ludlow, P. 1997. On the relevance of slow switching. *Analysis* 57:285-86.
- Ludlow, P. & Martin, N. 1998. *_Externalism and Self-Knowledge_*. CSLI.
- Macdonald, C. 1995. Externalism and first-person authority. *Synthese* 104:99-122.
On reconciling externalism with the non-evidential character of first-person knowledge.
- Macdonald, C. 1998. Externalism and authoritative self-knowledge. In (C. Wright, P. Smith, & C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.
- Macdonald, C., Smith, P. & Wright, C. 1998. *_Knowing Our Own Minds: Essays in Self-Knowledge_*. Oxford University Press.
- McKinsey, M. 1987. Apriorism in the philosophy of language. *Philosophical Studies* 52:1-32.
Argues that we can know the meaning of our words a priori. Analyzes twin earth cases by separating propositional meaning from linguistic meaning, which is indexical, fixes reference, and is knowable a priori.
- McKinsey, M. 1991. Anti-individualism and privileged access. *Analysis* 51:9-16.
Contra Burge: if there are conceptual connections between wide contents and the external world, then we can't know wide contents a priori, as otherwise we could know a priori that the world exists.
- McKinsey, M. 1994. Accepting the consequences of anti-individualism. *Analysis* 54:124-8.
Reply to Brueckner 1992: The claim that belief metaphysically necessitate external facts is trivial. Almost all states do that, for Kripkean reason.
- McLaughlin, B.P. & Tye, M. 1998. Externalism, Twin Earth, and self-knowledge. In (C. Macdonald, P. Smith, & C. Wright, eds) *_Knowing Our Own Minds: Essays in Self-Knowledge_*. Oxford University Press.
- McLaughlin, B.P., & Tye, M. 1998. Is content-externalism compatible with privileged access? *Philosophical Review* 107:349-380.
- Miller, R.W. 1997. Externalist self-knowledge and the scope of the a priori. *Analysis* 57:67-74.
- Peacocke, C. 1996. Entitlement, self-knowledge, and conceptual redeployment. *Proceedings of the Aristotelian Society* 96:117-58.
- Raffman, D. 1998. First-person authority and the internal reality of beliefs. In (C. Wright, B. Smith, & C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.
- Sawyer, S. 1998. Privileged access to the world. *Australasian Journal of Philosophy* 76:523-533.

Schiffer, S. 1992. Boghossian on externalism and inference. *Philosophical Issues* 2:29-38.

Szubka, T. 2000. Meaning rationalism, a priori, and transparency of content. *Philosophical Psychology* 13:491-503.

Tye, M. 1998. Externalism and memory. *Aristotelian Society Supplement* 72:77-94.

Warfield, T.A. 1992. Privileged self-knowledge and externalism are compatible. *Analysis* 52:232-37.

Boghossian's argument that externalism threatens self-knowledge fails: twin cases needn't be relevant alternatives (unless they are actual), so they don't threaten knowledge of content, by the usual standards of knowledge.

Warfield, T.A. 1995. Knowing the world and knowing our minds. *Philosophy and Phenomenological Research*.

Argues that externalism and self-knowledge imply the falsity of skepticism (though externalism alone does not). And arguments against externalist self-knowledge are no better than standard skeptical arguments.

Warfield, T.A. 1997. Externalism, privileged self-knowledge, and the irrelevance of slow switching. *Analysis* 57:282-84.

Wyler, T. 1994. First-person authority and singular thoughts. *Zeitschrift fur Philosophie Forschung* 48:585-94.

2.2g The Status of Narrow Content

Adams, F., Drebusenko, D., Fuller, G. & Stecker, R. 1990. Narrow content: Fodor's folly. *Mind and Language* 5:213-29.

Traces and criticizes Fodor's position on narrow content. Argues that narrow content isn't content, and doesn't explain behavior. Fun but arguable.

Adams, F. & Fuller, G. 1992. Names, contents, and causes. *Mind and Language* 7:205-21.

Argues that problems with names don't require an appeal to narrow content in explanation. Broad content plus associated descriptions will do the job.

Antony, L. 1989. Semantic anorexia: On the notion of content in cognitive science. In (G. Boolos, ed) *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.

Representational cognitive science has no need for narrow content -- wide contents and formal properties can do all the work. Argues that the semantics of mental expressions needn't mirror the semantics of language.

Aydede, M. 1997. Has Fodor really changed his mind on narrow content? *Mind and Language* 12:422-58.

Baker, L.R. 1985. A farewell to functionalism. *Philosophical Studies* 48:1-14.

Argues that type-identical functional states can differ in narrow content, so methodological solipsism fails. Uses the example of identical programs for playing chess and arms negotiations.

Baker, L.R. 1985. Just what do we have in mind? *Midwest Studies in Philosophy* 10:25-48.

Some implausible twin cases trying to show that mental life can vary wildly while preserving physical/computational state. Bizarre.

Baker, L.R. 1986. Content by courtesy. *Journal of Philosophy* 84:197-213.

- Baker, L.R. 1987. Saving Belief. Princeton University Press.
Lots of arguments against narrow content. Very stimulating, though wrong.
- Biro, J.I. 1992. In defense of social content. *Philosophical Studies* 67:277-93.
Contra Loar 1988, the contents of "that"-clauses often reflects psychological content, even if it sometimes does not. We don't need narrow content.
- Block, N. 1991. What narrow content is not. In (B. Loewer & G. Rey, eds) Meaning in Mind: Fodor and his Critics. Blackwell.
There are big problems specifying the "mapping" and the relevant contexts for Fodor's theory noncircularly. Narrow content either collapses into syntax or is too coarse-grained. Nontrivial narrow content must be holistic.
- Block, N. 1995. Ruritania revisited. In (E. Villanueva, ed) Content. Ridgeview.
- Brown, C. 1993. Belief states and narrow content. *Mind and Language* 8:343-67.
Criticizes the "bracketing" strategy of Stich and Walker, and argues that intrinsic belief state should be individuated according to how it embeds in different environments. With a comparison with Fodor's related theory.
- Chalmers, D.J. 1994. The components of content. Manuscript.
Argues for a two-dimensional intensional theory, with different kinds of intensions constituting notional and relational content. Notional content governs the dynamics of thought and behavior, and is primary in explanation.
- Davies, M. 1986. Externality, psychological explanation, and narrow content. *Proceedings of the Aristotelian Society* 60:263-83.
Comments on Fodor 1987. Fodor doesn't make a conclusive case against externalism; but narrow content may be promising, and inexpressibility doesn't pose any real problems. With comparisons to neo-Fregean theories.
- Dennett, D.C. 1983. Beyond belief. In (A. Woodfield, ed) Thought and Object. Oxford University Press. Reprinted in The Intentional Stance (MIT Press, 1987).
What matters are not propositional attitudes but notional attitudes; but it's hard to calibrate notional worlds. Very nice.
- Devitt, M. 1990. The narrow representational theory of mind. In (W. Lycan, ed) Mind and Cognition. Blackwell.
Not syntactic psychology nor wide psychology, but narrow psychology.
- Field, H. 1989. "Narrow" aspects of intentionality and the information-theoretic approach to content. In (E. Villanueva, ed) Information, Semantics, and Epistemology. Blackwell.
- Fodor, J.A. 1987. Individualism and supervenience. In Psychosemantics. MIT Press.
Science taxonomizes by causal powers, which are locally supervenient, so psychology needs a narrow notion of content. Proposes that a relativized notion -- a function from context to extension -- can do the job. Nice.
- Jackson, F., and Pettit, P. 1993. Some content is narrow. In (J. Heil and A. Mele, eds) Mental Causation. Oxford University Press.
Argues that folk psychology needs a notion of narrow content to provide robust predictive behavioral generalizations that covers doppelgangers. If not, then some behavioral patterns would be flukey.
- LePore, E. & Loewer, B. 1986. Solipsistic semantics. *Midwest Studies in Philosophy* 10:595-614.
There's no good way to construe narrow content. Phenomenologist strategy is intrinsically wide, and indexicalist strategy can't specify content.

LePore, E. & Loewer, B. 1989. Dual aspect semantics. In (S. Silvers, ed) _ReRepresentation_. Kluwer.

Loar, B. 1987. Social content and psychological content. In (R. Grimm & D. Merrill, eds) _Contents of Thought_. University of Arizona Press.
Uses examples to argue that psychological content is not fixed by the content of "that"-clauses in belief ascription, and vice versa. We require a subtler kind of narrow content to capture what's going on.

Loar, B. 1987. Subjective intentionality. *Philosophical Topics* 15:89-124.

Maloney, J.C. 1991. Saving psychological solipsism. *Philosophical Studies* 61:267-83.
Contests the "provoked/aggravated assault" example of Baker 1986. If they're doppelgangers, then their narrow content can't differ.

Manfredi, P. 1993. Two routes to narrow content: both dead ends. *Philosophical Psychology* 6:3-22.

McDermott, M. 1986. Narrow content. *Australasian Journal of Philosophy* 64:277-88.
Narrow beliefs are de re beliefs about our inputs and outputs.

McGilvray, J. 1998. Meanings are syntactically individuated and found in the head. *Mind and Language* 13:225-280.

Putnam, H. 1987. Fodor and Block on narrow content. In _Representation and Reality_. MIT Press.
Against perceptual-prototype and conceptual-role accounts of narrow content.

Quillen, K. 1986. Propositional attitudes and psychological explanation. *Mind and Language* 1:133-57.
Can't get a 'mode of presentation' account of narrow content to work, either through description theory or prototypes. Psych should be non-individualist.

Recanati, F. 1990. Externalism and narrow content. *Nous*.
There are levels of narrowness, varying by whether independence is of actual or normal environment. Argues that this can be consistent with externalism.

Recanati, F. 1994. How narrow is narrow content? *Dialectica* 48:209-29.

Schiffer, S. 1989. Fodor's character. In (E. Villanueva, ed) _Information, Semantics, and Epistemology_. Blackwell.

Segal, G. 2000. _A Slim Book about Narrow Content_. MIT Press.

Silverberg, A. 1995. Narrow content: A defense. *Southern Journal of Philosophy* 33:109-27.

Stalnaker, R.C. 1990. Narrow content. In (C.A. Anderson & J. Owens, eds) _Propositional Attitudes_. CSLI.
On some problems with narrow content, contra Loar 1987. Narrow content is hard to spell out with "diagonal" propositions. Loar doesn't show that psychological content is narrow. With some remarks on privileged access.

Stich, S.P. 1991. Narrow content meets fat syntax. In (B. Loewer & G. Rey, eds) _Meaning in Mind: Fodor and his Critics_. Blackwell.
Argues that narrow content is still too coarse-grained for explanation, classifying psychologically distinct states together. Use syntax instead.

Taylor, K. 1989. Supervenience and levels of meaning. *Southern Journal of Philosophy* 27:443-58.

Argues that the partial character construal of narrow content is not interestingly semantic. It collapses into syntax or phenomenology.

Taylor, K. 1989. Narrow content functionalism and the mind-body problem. *Nous* 23:355-72.

Uses a "fraternal twin earth" thought experiment to show that even de dicto attributions don't supervene on narrow role, and that narrow content can't be explicated descriptively unless it collapses into phenomenalism.

Vaughan, R. 1989. Searle's narrow content. *Ratio* 2:185-90.

White, S. 1982. Partial character and the language of thought. *Pacific Philosophical Quarterly* 63:347-65.

Replies to Burge/Stich arguments by introducing partial character -- a function from context to content, analogous to Kaplan's character -- as the semantic property determined by functional state and relevant to explanation.

White, S. 1992. Narrow content and narrow interpretation. In *The Unity of the Self*. MIT Press.

Argues for an account of narrow content in terms of notional worlds, by considering "objective optimality" across worlds. This allows for a sort of narrow radical interpretation. With arguments against Stalnaker.

Williams, M. 1990. Social norms and narrow content. *Midwest Studies in Philosophy* 15:425-462.

Narrow content theories can't handle the normativity of content. In-depth treatment of Burge cases and of the failures of causal and conceptual-role accounts. Normativity is fundamentally social. A long, interesting paper.

Williamson, T. 1998. The broadness of the mental: Some logical issues. *Philosophical Perspectives* 12:389-410.

2.2h Miscellaneous

Brook, D. 1992. Substantial mind. *South African Journal of Philosophy* 1:15-21.

Brown, D.J. 1993. Swampman of La Mancha. *Canadian Journal of Philosophy* 23:327-48.

An entertaining fable about a swampthing doppelganger of a murder witness. Does he have content? With plot twists about personal identity.

Brown, D.J. 1996. A furry tile about mental representation. *Philosophical Quarterly* 185:448-66.

Buekens, F. 1994. Externalism, content, and causal histories. *Dialectica* 48:267-86.

de Vries, W.A. 1996. Experience and the swamp creature. *Philosophical Studies* 82:55-80.

Argues that a swampthing isn't intelligent or intentional, with different physiological processes and no sensations, as these are functional kinds.

Edwards, S. 1994. *Externalism in the Philosophy of Mind*. Avebury.

Engel, P. 1987. Functionalism, belief, and content. In (Torrance, ed) *The Mind and the Machine*. Horwood.

Gauker, C. 1991. Mental content and the division of epistemic labour. *Australasian Journal of Philosophy* 69:302-18.

Gibbons, J. 1993. Identity without supervenience. *Philosophical Studies*

70:59-79.

Houghton, D. 1997. Mental content and external representations: internalism, anti-internalism. *Philosophical Quarterly* 47:159-77.

Jackson, F. & Pettit, P. 1988. Functionalism and broad content. *Mind* 97:318-400.

Should construe functionalism broadly rather than narrowly; then can handle the problem of broad content.

Katz, J. 1990. The domino theory. *Philosophical Studies* 58:3-39.
Anti-intensional arguments are not independent but a series of dominos.
Quine/Quine/Davidson/Putnam/Burge rise and fall together.

Macdonald, C. 1990. Weak externalism and mind-body identity. *Mind* 99:387-404.

McCulloch, G. 1995. *_The Mind and its World_*. Routledge.

McGinn, C. 1982. The structure of content. In (A. Woodfield, ed) *_Thought and Object_*. Oxford University Press.
Belief content has two distinct elements, one causal-explanatory, the other truth-related.

Owens, J. 1987. In defense of a different Doppelganger. *Philosophical Review* 96:521-54.

Owens, J. 1992. Psychophysical supervenience: Its epistemological foundation. *Synthese* 90:89-117.

Pereboom, D. 1995. Conceptual structure and the individuation of content. *Philosophical Perspectives* 9:401-428.

Preti, C. 2000. Belief and desire under the elms. *Protosociology* 14:270-284.

Rey, G. 1992. Semantic externalism and conceptual competence. *Proceedings of the Aristotelian Society* 66:315-33.
Supplements externalist "locking" theories of content with an account of internal "conceptions" by which thoughts lock onto environmental kinds, with that aid of dthat operators, thus solving various philosophical problems.

Rowlands, M. 1995. Externalism and token-token identity. *Philosophia* 24:359-75.

Rowlands, M. 1999. *_The Body in Mind: Understanding Cognitive Processes_*. Cambridge University Press.

Rudd, A. 1997. Two types of externalism. *Philosophical Quarterly* 47:501-7.

Seager, W.E. 1992. Externalism and token identity. *Philosophical Quarterly* 42:439-48.

Stalnaker, R.C. 1989. On what's in the head. *Philosophical Perspectives* 3:287-319.

Thomas, J. 1996. Analogies and the mind of the replica: Sunburn, the little green bug, and the fake plant. *Philosophical Quarterly* 46:364-371.

Walker, V. 1990. In defense of a different taxonomy: A reply to Owens. *Philosophical Review* 99.
Contra Owens 1987: wide intentional descriptions and molar bodily descriptions don't exhaust the options. A bracketing strategy gives a narrow intentional taxonomy of mental states.

Williams, M. 1990. Externalism and the philosophy of mind. *Philosophical*

Quarterly 40:352-80.

Woodfield, A. 1986. Two categories of content. *Mind and Language* 1:319-54.

2.3 Causal Theories of Content

2.3a Information-Based Accounts (Dretske, etc)

Barwise, J. & Perry, J. 1983. *Situations and Attitudes*. MIT Press.

Barwise, J. 1986. Information and circumstance. *Notre Dame Journal of Formal Logic*.

Defending information against Fodor 1986. Information is objective but relational, and depends on the relevant constraints between representation and environment. Circumstances play a vital role.

Barwise, J. 1987. Unburdening the language of thought. *Mind and Language*.

Bogdan, R.J. 1988. Information and semantic cognition: An ontological account. *Mind and Language*.

From material (formal) info to semantic info via teleology; from semantic information to representation via internal structure. Cute. With a good reply by Israel, and a terse reply by Dretske.

Bogdan, R.J. 1987. Mind, content and information. *Synthese*.

Clark, A. 1993. Mice, shrews, and misrepresentation. *Journal of Philosophy* 90:290-310.

Uses information theory to analyze misrepresentation. A signal represents what it carries most information about, not what it correlates best with. Treating some signals as noise can increase information content.

Coulter, J. 1995. The informed neuron: Issues in the use of information theory in the behavioral sciences. *Minds and Machines* 5:583-96.

Dretske, F. 1981. *Knowledge and the Flow of Information*. MIT Press.

Defines knowledge content is in terms of information-flow from events, and applies to various aspects of psychology.

Dretske, F. 1983. *Precis of Knowledge and the Flow of Information*. Behavioral and Brain Sciences 6:55-90.

A summary of the book, with commentary and replies.

Dretske, F. 1990. Putting information to work. In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

On the causal role of information (as opposed to meaning). Information is causally efficacious if considered with respect to learning. With commentary by Brian Smith.

Fodor, J.A. 1986. Information and association. *Notre Dame Journal of Formal Logic* 27.

Transmission of information is no good without the encoding of information. With criticisms of associative networks, which transmit without encoding, and criticism of Barwise & Perry's account of attunement to a relation.

Fodor, J.A. 1987. A situated grandmother. *Mind and Language*.

Foley, R. 1987. Dretske's 'information-theoretic' account of knowledge. *Synthese*.

Gjelsvik, O. 1991. Dretske on knowledge and content. *Synthese* 86:425-41.

Grandy, R. 1987. Information-based epistemology, ecological epistemology and epistemology naturalized. *Synthese* 70:191-203.

Shannon's notion of information is more useful for naturalized epistemology than Dretske's.

Hardcastle, V.G. 1994. Indicator semantics and Dretske's function. *Philosophical Psychology* 7:367-82.

Heller, M. 1991. Indication and what might have been. *Analysis* 51:187-91.

We need to analyze indication in terms of "close enough" worlds; the relevant conditionals are "might"-conditionals.

Israel, D. & Perry, J. 1990. What is information? In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

Jackendoff, R. 1985. Information is in the mind of the beholder. *Linguistics and Philosophy* 8:23-33.

Argues that a representationalist theory of semantics beats a realist one.

Loewer, B. 1987. From information to intentionality. *Synthese*.

Morris, W.E. 1990. The regularity theory of information. *Synthese* 82:375-398.

Dretske has problems with ruling out alternative possibilities; and there is a gap between information-caused belief and knowledge.

Savitt, S. 1987. Absolute informational content. *Synthese* 70:185-90.

Makes a distinction between absolute information and information that's relative to other knowledge.

Sayre, K.M. 1986. Intentionality and information processing: An alternative model for cognitive science. *Behavioral and Brain Sciences* 9:121-38.

Sayre, K.M. 1987. Cognitive science and the problem of semantic content. *Synthese* 70:247-69.

On problems with a computational approach to content: computers process info(t), the non-semantic content of communication theory, not info(s), or semantic content. Get info(s) from efficient processing of mutual info(t).

Sturdee, D. 1997. The semantic shuffle: Shifting emphasis in Dretske's account of representational content. *Erkenntnis* 47:89-104.

Taylor, K. 1987. Belief, information and semantic content: A naturalist's lament. *Synthese* 71:97-124.

Usher, M. 2001. A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language* 16:331-334.

Winograd, T. 1987. Cognition, attunement and modularity. *Mind and Language*.

Zalabardo, J.L. 1995. A problem for information-theoretic semantics. *Synthese* 105:1-29.

2.3b Asymmetric Dependence (Fodor)

Fodor, J.A. 1987. Meaning and the world order. In *Psychosemantics*. MIT Press.

Defends and refines a causal theory, using the notion of asymmetric dependence of a token upon the world.

Fodor, J.A. 1990. A theory of content II. In *A Theory of Content*. MIT Press.

Defending the asymmetric dependence theory against various objections.

- Adams, F. & Aizawa, K. 1992. 'X' means X: Semantics Fodor-style. *Minds and Machines* 2:175-83.
- Adams, F. & Aizawa, K. 1993. Fodorian semantics, pathologies, and "Block's problem". *Minds and Machines* 3:97-104.
- Adams, F. & Aizawa, K. 1994. 'X' means X: Fodor/Warfield semantics. *Minds and Machines* 4:215-31.
- Adams, F. & Aizawa, K. 1997. Fodor's asymmetric causal dependency theory and proximal projections. *Southern Journal of Philosophy* 35:433-437.
- Antony, L. & Levine, J. 1991. The nomic and the robust. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.
- Baker, L.R. 1990. On a causal theory of content. *Philosophical Perspectives*.
- Baker, L.R. 1991. Has content been naturalized? In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.
- Bernier, P. 1993. Narrow content, context of thought, and asymmetric dependence. *Mind and Language* 8:327-42.
- Boghossian, P. 1991. Naturalizing content. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.
Argues that Fodor's theory is a type-1 theory, requiring naturalistically specifiable circumstances in which a symbol is only caused by its referent; and that these theories fail for various reasons, e.g. verificationism.
- Cram, H-R. 1992. Fodor's causal theory of representation. *Philosophical Quarterly* 42:56-70.
Fodor's theory has counterexamples and can't explain its counterfactuals; but we can save it by borrowing from Dretske's account of misrepresentation.
- Gibson, M. 1996. Asymmetric dependencies, ideal conditions, and meaning. *Philosophical Psychology* 9:235-59.
- Loar, B. 1991. Can we explain intentionality? In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.
- Maloney, J.C. 1990. Mental misrepresentation. *Philosophy of Science* 57:445-58.
- Manfredi, P.A. & Summerfield, D.M. 1992. Robustness without asymmetry: A flaw in Fodor's theory of content. *Philosophical Studies* 66:261-83.
- Rupert, R. 2000. Dispositions indisposed: Semantic atomism and Fodor's theory of content. *Pacific Philosophical Quarterly* 81:325-349.
- Seager, W.E. 1993. Fodor's theory of content: problems and objections. *Philosophy of Science* 60:262-77.
- Wallis, C. 1995. Asymmetric dependence, representation, and cognitive science. *Southern Journal of Philosophy* 33:373-401.
- Warfield, T.A. 1994. Fodorian semantics: A reply to Adams and Aizawa. *Minds and Machines* 4:205-14.

2.3c Causal Accounts, General

- Aizawa, K. 1994. Lloyd's dialectical theory of representation. *Mind and Language* 9:1-24.

Cummins, R. 1989. Representation and covariation. In (S. Silvers, ed) _ReRepresentation_. Kluwer.

Cummins, R. 1997. The LOT of the causal theory of mental content. *Journal of Philosophy* 94:535-542.

Fodor, J.A. 1984. Semantics, Wisconsin style. *Synthese* 59:231-50. Reprinted in _RePresentations_ (MIT Press, 1980).

A somewhat sympathetic commentary on the Dretske/Stampe causal theories, but raising the problem of misrepresentation.

Fodor, J.A. 1990. Information and representation. In (P. Hanson, ed) _Information, Language and Cognition_. University of British Columbia Press.

Godfrey-Smith, P. 1989. Misinformation. *Canadian Journal of Philosophy* 19:533-50.

On various attempts to solve the error problem and why they don't work.

Godfrey-Smith, P. 1991. Signal, decision, action. *Journal of Philosophy* 88:709-22.

World-head reliability is just as important as head-world reliability. With arguments and examples from signal detection theory.

Jacquette, D. 1996. Lloyd on intrinsic natural representation in simple mechanical minds. *Minds and Machines* 6:47-60.

Maloney, J.C. 1994. Content: Covariation, control, and contingency. *Synthese* 100:241-90.

McLaughlin, B.P. 1987. What is wrong with correlational psychosemantics. *Synthese*.

Ray, G. 1997. Fodor and the inscrutability problem. *Mind and Language* 12:475-89.

Stampe, D. 1977. Towards a causal theory of linguistic representation. *Midwest Studies in Philosophy* 2:42-63.

Stampe, D. 1986. Verificationism and a causal account of meaning. *Synthese* 69:107-37.

Stampe, D. 1991. Content, context, and explanation. In (E. Villanueva, ed) _Information, Semantics, and Epistemology_. Blackwell.

Viger, C.D. 2001. Locking on to the language of thought. *Philosophical Psychology* 14:203-215.

Warmbrod, K. 1992. Primitive representation and misrepresentation. *Topoi* 11:89-101.

Weitzman, L. 1996. What makes a causal theory of content anti-skeptical? *Philosophy and Phenomenological Research* 56:299-318.

2.3d Teleological Approaches (Millikan, etc)

Adams, F. & Aizawa, K. 1997. Rock beats scissors: Historicalism fights back. *Analysis* 57:273-81.

Agar, N. 1993. What do frogs really believe? *Australasian Journal of Philosophy* 71:1-12.

Argues that a teleological account can resolve content indeterminacies, by an appeal to counterfactuals in examining what properties were selected for.

Bogdan, R. 1994. *_Grounds for Cognition: How Goal-Guided Behavior Shapes the Mind_*. Lawrence Erlbaum.

Braddon-Mitchell, D. & Jackson, F. 1997. The teleological theory of content. *Australasian Journal of Philosophy* 75:474-89.

Clarke, M. 1996. Darwinian algorithms and indexical representation. *Philosophy of Science* 63:27-48.

Dennett, D.C. 1988. Fear of Darwin's optimizing rationale. Manuscript.
Defends evolutionary theories of content against Fodor.

Dennett, D.C. 1988. Evolution, error and intentionality. In *_The Intentional Stance_*. MIT Press.
Attacks original intentionality (Fodor/Burge/Dretske/Searle/Kripke) -- our intentionality, if anything, is derived through evolution, and so is as indeterminate as that of an artifact.

Dretske, F. 1986. Misrepresentation. In (R. Bogdan, ed) *_Belief: Form, Content, and Function_*. Oxford University Press.
Tries to deal with misrepresentation by appealing to function.

Elder, C.L. 1998. What versus how in naturally selected representations. *Mind* 107:349-363.

Fodor, J.A. 1990. Psychosemantics, or, Where do truth conditions come from? In (W. Lycan, ed) *_Mind and Cognition_*. Blackwell.
Truth conditions are "entry conditions" for belief under "normal function". Later repudiated.

Fodor, J.A. 1990. A theory of content I. In *_A Theory of Content_*. MIT Press.
Teleological solutions can't work, because of underdetermination and so on.

Keeley, B. 1999. Fixing content and function in neurobiological systems: The neuroethology of electroreception. *Biology and Philosophy* 14:395-430.

Lalor, B.J. 1998. Swampman, etiology, and content. *Southern Journal of Philosophy* 36:215-232.

Levine, J. 1996. Swampjoe: mind or simulation? *Mind and Language* 11:86-91.

Macdonald, G. 1989. Biology and representation. *Mind and Language* 4:186-200.

Matthen, M. 1988. Biological functions and perceptual content. *Journal of Philosophy* 85:5-27.

Millikan, R.G. 1979. An evolutionist approach to language. *Philosophy Research Archives* 5.

Millikan, R.G. 1984. *_Language, Thought and Other Biological Categories_*. MIT Press.
An evolutionary account of thought, content, and various intentional phenomena, appealing to proper functions and adaptational role to individuate contents.

Millikan, R.G. 1986. Thoughts without laws: Cognitive science with content. *Philosophical Review* 95:47-80.
The content of a desire is its adaptational Proper Function; the content of a belief is its Normal Condition for success.

Millikan, R.G. 1989. Biosemantics. *Journal of Philosophy* 86:281-97.
Representation content is determined by the consumption of a representation,

not its production. The representation-world correspondence is best taken as a normal condition for the consumer's function.

Millikan, R.G. 1989. In defense of proper functions. *Philosophy of Science* 56:288-302.

Millikan, R.G. 1990. Compare and contrast Dretske, Fodor, and Millikan on teleosemantics. *Philosophical Topics* 18:151-61.
Contrasting positions on the role of representation production and consumption; also on the role of reliability, articulateness, and learning.

Millikan, R.G. 1991. Speaking up for Darwin. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.
A reply to some of Fodor's criticisms of teleological theories in *Psychosemantics* and elsewhere. With some remarks on Fodor's asymmetric dependence theory.

Millikan, R.G. 1993. *White Queen Psychology and Other Essays for Alice*. MIT Press,
A collection of papers on teleological semantics and other issues about psychology and mental content.

Millikan, R.G. 1996. On swampkinds. *Mind and Language* 11:103-17.

Millikan, R.G. 1997. Troubles with Wagner's reading of Millikan. *Philosophical Studies* 86:93-96.

Neander, K. 1995. Misrepresenting and malfunctioning. *Philosophical Studies* 79:109-41.

Neander, K. 1995. Dretske's innate modesty. *Australasian Journal of Philosophy* 74:258-74.

Neander, K. 1996. Swampman meets swampcow. *Mind and Language* 11:118-29.
It's not unreasonable to deny a swampthing beliefs: swampcows aren't cows and swamphearts aren't hearts. Semantic norms are plausibly grounded in biological norms and so in history.

Newton, N. 1992. Dennett on intrinsic intentionality. *Analysis* 52:18-23.
Contra Dennett 1988, designed creatures can have intrinsic (if not original) intentionality. Overall purpose is dependent on designer's goals, but specific contents need not be.

Papineau, D. 1984. Representation and explanation. *Philosophy of Science* 51:550-72.
A teleological theory of belief/desire contents: the satisfaction conditions for a desire are those effects for which it was selected; truth conditions for a belief are circumstances resulting in satisfaction of desires.

Papineau, D. 1990. Truth and teleology. In (D. Knowles, ed) *Explanation and its Limits*. Cambridge University Press.
Best theory is combination of a success-guaranteeing account of truth-conditions with a teleological account of desire.

Papineau, D. 1991. Teleology and mental states. *Proceedings of the Aristotelian Society* 65.

Papineau, D. 1996. Doubtful intuitions. *Mind and Language* 11:130-32.

Papineau, D. 1998. Teleosemantics and indeterminacy. *Australasian Journal of Philosophy* 76:1-14.

Papineau, D. 2001. The status of teleosemantics, or how to stop worrying about

Swampman. Australasian Journal of Philosophy 79:279-89.

Pickles, D. 1989. Intentionality, representation, and function. Sussex University, Cognitive Science Research Paper 140.

Combining the analysis-relative and historical accounts of function, and using these to give an account of intentionality: representation are produced by conditional productive functions. Argues against Fodor on indeterminacy.

Pietrowski, P.M. 1992. Intentionality and teleological error. Pacific Philosophical Quarterly 73:267-82.

Millikan's theory has an implausible consequence: creatures' belief contents can involve properties which they cannot discriminate. With examples.

Ross, D. & Zawidzki, T. 1994. Information and teleosemantics. Southern Journal of Philosophy 32:393-419.

Rountree, J. 1997. The plausibility of teleological content ascriptions: A reply to Pietroski. Pacific Philosophical Quarterly 78:404-??.

Rowlands, M. 1996. Teleological semantics. Mind 106:279-304.

Rupert, R.D. 1999. Mental representations and Millikan's theory of intentional content: Does biology chase causality? Southern Journal of Philosophy 37:113-140.

Sehon, S.R. 1994. Teleology and the nature of mental states. American Philosophical Quarterly 31:63-72.

Shapiro, L. 1996. Representation from bottom to top. Canadian Journal of Philosophy 26:523-42.

Shapiro, L. 1992. Darwin and disjunction: Foraging theory and univocal assignments of content. Philosophy of Science Association 1992, 1:469-80.

Sullivan, S.R. 1993. From natural function to indeterminate content. Philosophical Studies 69:129-37.

Wagner, S. 1996. Teleosemantics and the troubles of naturalism. Philosophical Studies 82:81-110.

Teleosemantics has big problems with indeterminacy, holism, false belief, and "psychophysical normalcy" in causation. So do all naturalistic stories.

Zawidzki, T. & Ross, D. 1994. Information and teleosemantics. Southern Journal of Philosophy 32:393-419.

2.3e Conceptual Role Approaches

Block, N. 1986. Advertisement for a semantics for psychology. Midwest Studies in Philosophy 10:615-78.

An in-depth program for conceptual-role semantics, and its role in a two-factor account of meaning. Also a defense of narrow content.

Block, N. 1988. Functional role and truth conditions. Proceedings of the Aristotelian Society 61:157-181.

A defense of functional role semantics, and an account of its relation to truth-conditional factors. A two-factor theory will handle wide content.

Boghossian, P.A. 1994. Inferential-role semantics and the analytic/synthetic distinction. Philosophical Studies.

No matter how we understand the denial of the analytic/synthetic distinction, the falsity of inferential-role semantics does not follow. The meaning-constitutive inferences needn't be the analytic inferences.

Brandom, R. 1994. *Making It Explicit*. Harvard University Press.

Brandom, R. 1994. Reasoning and representing. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind*. Kluwer.

Cummins, R. 1992. Conceptual role semantics and the explanatory role of content. *Philosophical Studies* 65:103-127.

CRS conflates representation content and attitude content (which depends on a representation's "target"), so can't handle representation content; it makes all content-based explanations vacuous; and it can't handle error properly.

Field, H. 1977. Logic, meaning, and conceptual role. *Journal of Philosophy* 74:379-409.

Explicates conceptual role in terms of conditional probability, and analyzes meaning as conceptual role plus reference. With remarks on truth, descriptions, and synonymy.

Field, H. 1978. Mental representation. *Erkenntnis* 13:9-61.

Fodor, J.A. & LePore, E. 1991. Why meaning (probably) isn't conceptual role. *Mind and Language* 6:328-43.

Conceptual role semantics isn't compatible with compositional semantics and the denial of an analytic/synthetic distinction, as full conceptual roles aren't compositional, and there's no way to specify a relevant subset.

Harman, G. 1974. Meaning and semantics. In (M. Munitz & P. Unger, eds) *Semantics and Philosophy*. New York University Press.

Harman, G. 1975. Language, thought, and communication. In (K. Gunderson, ed) *Language, Mind, and Knowledge*. University of Minnesota Press.

Harman, G. 1982. Conceptual role semantics. *Notre Dame Journal of Formal Logic* 28:242-56.

Meaning and content is determined by the role of symbols in thought (e.g. inference and perception). With remarks on indeterminacy, context-dependence, the linguistic division of labor, qualia, speech acts, and more.

Horowitz, A. 1992. Functional role and intentionality. *Theoria* 58:197-218.

Loar, B. 1982. Conceptual role and truth conditions. *Notre Dame Journal of Formal Logic* 23:272-83.

On the relation between conceptual role and truth-conditions. Contra Harman, truth-conditions are to an extent independent of conceptual role.

Loewer B. 1982. The role of 'Conceptual role semantics'. *Notre Dame Journal of Formal Logic* 23:305-15.

Contra Harman 1982, truth-conditions are central to a semantic theory.

Perlman, M. 1997. The trouble with two-factor conceptual role theories. *Minds and Machines* 7:495-513.

Silverberg, A. 1992. Putnam on functionalism. *Philosophical Studies* 67:111-31.

Argues against Putnam 1987 that conceptual role plays an important role in determining meaning. Appeals to the induction theory of Holland et al.

Toribio, J. 1997. Twin pleas: Probing content and compositionality. *Philosophy and Phenomenological Research* 57:871-89.

Warfield, T.A. 1993. On a semantic argument against conceptual role semantics. *Analysis* 53:298-304.

Contra Fodor and Lepore, meanings can be compositional even if inferential

roles are not, as long as meanings only supervene on inferential role.

2.3f Theories of Content, Misc

Bestor, T.W. 1991. Naturalizing semantics: New insights or old folly? *Inquiry* 34:285-310.

Brook, A. & Stainton, R. 1997. Fodor's new theory of content and computation. *Mind and Language* 12:459-74.

Callaway, H.G. 1995. Intentionality naturalized: Continuity, reconstruction, and instrumentalism. *Dialectica* 49:147-68.

Churchland, P.M. & Churchland, P.S. 1983. Stalking the wild epistemic engine. *Nous* 17:5-18.

On "translational" (conceptual) and "calibrational" (referential) content.
Relation of content issues to computational issues.

Cummins, R. 1989. *_Meaning and Mental Representation_*. MIT Press.
Critiques other views, offers interpretational semantics.

Cummins, R. 1996. *_Representations, Targets, and Attitudes_*. MIT Press.

Dennett, D.C. 1991. Ways of establishing harmony. In (B. McLaughlin, ed) *_Dretske and his Critics_*. Blackwell.

On the ways in which meanings can come to cohere with their causal roles:
learning, natural selection, and design. Criticizes Dretske for undervaluing
the latter two: all three are in the same boat.

Dretske, F. 1986. Aspects of cognitive representation. In (M. Brand & R. Harnish, eds) *_The Representation of Knowledge and Belief_*. University of Arizona Press.

On the reference and content of representations. Reference is determined by
causation; content, i.e. representation "as", is determined by functional
role, when functioning normally in natural habitat.

Shapiro, L.A. 1997. The nature of nature: Rethinking naturalistic theories of intentionality. *Philosophical Psychology* 10:309-322.

Silvers, S. 1991. On naturalizing the semantics of mental representation. *British Journal for the Philosophy of Science* 42:49-73.

Stalnaker, R. 1991. How to do semantics for the language of thought. In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and His Critics_*. Blackwell.

On some tensions in Fodor's view of content: e.g. narrow content must be
dependent on functional role, which seems to lead to holism. The role of
denotational semantics as a defense is unclear.

2.4 Representation (General) [see also 4.2]

Bickhard, M. 1993. Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence* 5:285-33.

Blachowicz, J. 1997. Analog representation beyond mental imagery. *Journal of Philosophy* 94:55-84.

Chomsky, N. 1980. Rules and representations. *Behavioral and Brain Sciences* 3:1-61.

Clark, A. 1995. Moving minds: Situating content in the service of real-time success. *Philosophical Perspectives* 9:89-104.

Dalenoort, G.J. 1990. Toward a general theory of representation. *Psychological Research* 52:229-237.

Fodor, J.A. 1986. Why paramecia don't have mental representations. *Midwest Studies in Philosophy* 10:3-23.
Because paramecia can't respond to non-nomic properties of the stimulus.
Perceptual categories vs. sensory manifolds.

Freeman, W. & Skarda, C.A. 1990. Representations: who needs them? In (J. McGaugh, J. Weinberger, & G. Lynch) *_Brain Organization and Memory_*. Guilford Press.

Gillett, G. 1989. Representations and cognitive science. *Inquiry* 32:261-77.

Goldman, A. 1986. Constraints on representation. In (M. Brand & R. Harnish, eds) *_The Representation of Knowledge and Belief_*. University of Arizona Press.

Grush, R. 1997. The architecture of representation. *Philosophical Psychology* 10:5-23.

Hatfield, G. 1989. Computation, representation and content in noncognitive theories of perception. In (S. Silvers, ed) *_ReRepresentation_*. Kluwer.

Hogan, M. 1994. What is wrong with an atomistic account of mental representation. *Synthese* 100:307-27.

Jackendoff, R. 1991. The problem of reality. *Nous* 25:411-33.
On the philosophical (inward-out) vs. psychological (outward-in) approaches to the mind-world relation; the psychological approach is more useful in understanding representation. Internal reality is an imperfect construction.

Kukla, R. 1992. Cognitive models and representation. *British Journal for the Philosophy of Science* 43:219-32.

Lloyd, D. 1987. Mental representation from the bottom up. *Synthese* 70:23-78.

Lycan, W.G. 1989. Ideas of representation. In (Weissbord, ed) *_Mind, Value and Culture: Essays in Honor of E.M. Adams_*. Ridgeview.

Matthews, R.J. 1984. Troubles with representationalism. *Social Research* 51:1065-97.

Millikan, R.G. 1995. Pushmi-pullyu representations. *Philosophical Perspectives* 9:185-200.

Richardson, R.C. 1981. Internal representation: Prologue to a theory of intentionality. *Philosophical Topics* 12:171-212.

Shanon, B. 1991. Representations -- senses and reasons. *Philosophical Psychology* 4:355-74.
On different senses of "representation" -- external, experiential, mental locus, substrate of meaning, mediating functions, technicalpsychological.

Shanon, B. 1993. *_The Representational and the Presentational: An Essay on Cognition and the Study of Mind_*. Prentice-Hall.

Sober, E. 1976. Mental representations. *Synthese* 33:101-48.

Sterelny, K. 1995. Basic minds. *Philosophical Perspectives* 9:251-70.

van Gulick, R. 1982. Mental representation: A functionalist view. *Pacific Philosophical Quarterly* 63:3-20.

On the distinction between representation and representation-use.

Wallis, C. 1994. Representation and the imperfect ideal. *Philosophy of Science* 61:407-28.

2.5 The Explanatory Role of Content (Dretske, etc)

Adams, F. 1991. Causal contents. In (B. McLaughlin, ed) *_Dretske and his Critics_*. Blackwell.

On Dretske's account of the causal role of content. Addresses some objections: Dennett's worries about intrinsic intentionality, Fodor's about external causal powers, and some worries about syntax.

Baker, L.R. 1991. Dretske on the explanatory role of belief. *Philosophical Studies* 63:99-111.

Bogdan, R.J. 1989. Does semantics run the psyche? *Philosophy and Phenomenological Research* 49:687-700.

A critique of Fodor. Semantics per se doesn't cause. Also, Fodor's is an account of the what, not the how, of semantics. Somewhat bizarre.

Cummins, R. 1991. Mental meaning in psychological explanation. In (B. McLaughlin, ed) *_Dretske and his Critics_*. Blackwell.

Criticizes Dretske's account of the role of content, especially because of its dependence on an organism's history; also, it may not cohere with work in cognitive science. Argues for an interpretational, not a causal account.

Devitt, M. 1991. Why Fodor can't have it both ways. In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell.

Dretske, F. 1987. The explanatory role of content. In (R. Grimm & D. Merrill, eds) *_Contents of Thought_*. University of Arizona Press.

Content must explain why (not how) an internal state caused a certain output. The explanation is given in terms of what a state has historically indicated. With thermostats and sea-snails as examples. Comments by Cummins, and reply.

Dretske, F. 1988. *_Explaining Behavior: Reasons in a World of Causes_*. MIT Press.

Dretske, F. 1990. Does meaning matter? In (E. Villanueva, ed) *_Information, Semantics, and Epistemology_*. Blackwell.

Dretske, F. 1994. Reply to Slater and Garcia-Carpintero. *Mind and Language* 9:203-8.

Dretske, F. 1995. Reply: Causal relevance and explanatory exclusion. In (E. Villanueva, ed) *_Information, Semantics, and Epistemology_*. Blackwell.

Dretske, F. 1996. The explanatory role of content: Reply to Melnyk and Noordhof. *Mind and Language* 11:223-29.

Elder, C.L. 1996. Content and the subtle extensionality of "... explains ...". *Philosophical Quarterly* 46:320-32.

Fodor, J.A. 1986. Banish DisContent. In (J. Butterfield, ed) *_Language, Mind, and Logic_*. Cambridge University Press. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Garcia-Carpintero, M. 1994. Dretske on the causal efficacy of meaning. *Mind and Language* 9:181-202.

Hassrick, B. 1995. Fred Dretske on the explanatory role of semantic content.

Conference 6:59-66.

Horgan, T. 1991. Actions, reasons, and the explanatory role of content. In (B. McLaughlin, ed) *_Dretske and his Critics_*. Blackwell.
Distinguishes three problems of mental causation (extrinsic factors, exclusion of the nonphysical, anomalism). Criticizes Dretske's theory (can't handle unlearned or here-and-now reasons), offers a counterfactual account.

Melnyk, A. 1996. The prospects for Dretske's account of the explanatory role of belief. *Mind and Language* 11:203-15.

Noordhof, P. 1996. Accidental associations, local potency, and a dilemma for Dretske. *Mind and Language* 11:216-22.

Perry, J. & Israel, D. 1991. Fodor and psychological explanation. In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell.

Pylyshyn, Z.W. 1987. What's in a mind? *Synthese* 70:97-122.
We must individuate mental states by semantics, not just by function, as we need representation to capture generalizations about behavior; particularly due to the information-sensitivity and stimulus-independence of behavior.

Slater, C. 1994. Discrimination without indication: Why Dretske can't lean on learning. *Mind and Language* 9:163-80.

Wallis, C. 1994. Using representation to explain. In (E. Dietrich, ed) *_Thinking Computers and Virtual Persons_*. Academic Press.

2.6 Concepts -----

Barber, A. 1998. The pleonasticity of talk about concepts. *Philosophical Studies* 89:53-86.

Brown, H. 1986. Sellars, concepts, and conceptual change. *Synthese* 68:275-307.

Burge, T. 1993. Concepts, definitions, and meaning. *Metaphilosophy* 24:309-25.

Churchland, P.M. 1998. Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy* 95:5-32.

Cussins, A. 1990. The connectionist construction of concepts. In (M. Boden, ed) *_The Philosophy of AI_*. Oxford University Press.
Connectionism builds up concepts from the nonconceptual level. From nonconceptual content (e.g. perceptual experiences) to the emergence of objectivity.

Fodor, J. & Lepore, E. 1996. The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition* 58:253-70.

Fodor, J. 1995. Concepts: A potboiler. *Cognition* 50:133-51. Also in (E. Villanueva, ed) *_Content_*. Ridgeview.

Fodor, J. 1998. *_Concepts: Where Cognitive Science Went Wrong_*. Oxford University Press.

Franks, B. 1992. Realism and folk psychology in the ascription of concepts. *Philosophical Psychology* 5:369-90.

Gardenfors, P. 1997. Meanings as conceptual structures. In (M. Carrier & P. Machamer, eds) *_Mindsapes: Philosophy, Science, and the Mind_*. Pittsburgh University Press.

Gauker, C. 1993. An extraterrestrial perspective on conceptual development. *Mind and Language* 8:105-30.

Grandy, R.E. 1989. Concepts, prototypes, and information. In (E. Villanueva, ed) *_Information, Semantics, and Epistemology_*. Blackwell.

Jackendoff, R. 1989. What is a concept, that a person may grasp it? *Mind and Language* 4:68-102.

Khalidi, M.A. 1995. Two concepts of concept. *Mind and Language* 10:402-22.

Levine, A. & Bickhard, MH. 1999. Concepts: Where Fodor went wrong. *Philosophical Psychology* 12:5-23.

Livingston, K.R. 1989. Concepts, categories, and epistemology. *Philosophia* 19:265-300.

Neisser, U. (ed) 1981. *_Concepts and Conceptual Development_*. Cambridge University Press.

Osherson, D.N. & Smith, E.E. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition* 9:35-58.

Margolis, E. 1995. The significance of the theory analogy in the psychological study of concepts. *Mind and Language* 10:45-71.

Margolis, E. 1998. How to acquire a concept. *Mind and Language* 13:347-369.

Margolis, E. 1999. What is conceptual glue? *Minds and Machines* 9:241-255.

Margolis, E. & Laurence, S. 1999. *_Concepts: Core Readings_*. MIT Press.

Millikan, R.G. 1994. On unclear and indistinct ideas. *Philosophical Perspectives* 8:75-100.

Millikan, R.G. 1997. A common structure for concepts of individuals, stuffs, and kinds: More mama, more milk, and more mouse. *Behavioral and Brain Sciences*.

Peacocke, C. 1989. What are concepts? *Midwest Studies of Philosophy* 14.

Peacocke, C. 1989. Possession conditions: A focal point for theories of concepts. *Mind and Language* 4:51-56.

Peacocke, C. 1991. The metaphysics of concepts. *Mind* 100:525-46.

Peacocke, C. 1992. *_A Study of Concepts_*. MIT Press.

Peacocke, C. 1996. *Precis of _A Study of Concepts_*. *Philosophy and Phenomenological Research* 56:407-52.

A symposium on the book, with comments by Heal, Rey, Papineau.

Peacocke, C. 1996. Can a theory of concepts explain the a priori: A reply to Skorupski. *International Journal of Philosophical Studies* 4:154-60.

Peacocke, C. 1996. The relation between philosophical and psychological theories of concepts. In (P. Millican & A. Clark, eds) *_Machines and Thought_*. Oxford University Press.

Peacocke, C. 2000. Theories of concepts: A wider task. *European Journal of Philosophy* 8:298-321.

Perlman, M. 2000. *_Conceptual Flux: Mental Representation, Misrepresentation, and Concept Change_*. Kluwer.

- Pitt, D. 1999. In defense of definitions. *Philosophical Psychology* 12:139-156.
- Ramsey, W. 1992. Prototypes and conceptual analysis. *Topoi* 11:59-70.
On the significance of psychological work on concepts for philosophical conceptual analysis -- simple, precise analyses do not exist in general.
- Rey, G. 1983. Concepts and stereotypes. *Cognition* 15:237-62.
- Rips, L.J. 1995. The current status of research on concept combination. *Mind and Language* 10:72-104.
- Sellars, W. 1948. Concepts as involving laws and inconceivable without them. *Philosophy of Science* 15:287-313.
- Sellars, W. 1974. Conceptual change. In *Essays in Philosophy and its History*. Reidel.
- Smith, E.E. & Medin, D.L. 1981. *Categories and Concepts*. Harvard University Press.
- Sosa, E. 1993. Abilities, concepts, and externalism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.
On concepts as abilities, and on construals of abilities that lead to internalism and externalism. Maybe the relevant abilities are characterized externally but determined internally. Remarks on Putnam, Davidson, Burge.
- Thagard, P. 1990. Concepts and conceptual change. *Synthese* 82:255-74.
- van Brakel, J. 1991. Meaning, prototypes, and the future of cognitive science. *Minds and Machines* 1:233-57.
- Vision, G. 2001. Flash! Fodor splits the atom. *Analysis* 61:5-10.
- Weitz, M. 1988. *Theories of Concepts: A History of the Major Philosophical Traditions*. Routledge.
- Woodfield, A. 1991. Conceptions. *Mind* 100:547-72.

2.7 Meaning Holism

- Abbott, B. 2000. Fodor and Lepore on meaning similarity and compositionality. *Journal of Philosophy* 97:454-6.
- Becker, K. 1998. On the perfectly general nature of instability in meaning holism. *Journal of Philosophy* 95:635-640.
- Bilgrami, A. 1998. Why holism is harmless and necessary. *Philosophical Perspectives* 12:105-126.
- Block, N. 1995. An argument for holism. *Proceedings of the Aristotelian Society* 95:151-70.
Uses Putnam's "Ruritania" example to argue that narrow content, if it exists, is holistic. Twins in different communities start with same narrow content, diverge by acquiring new beliefs; so belief change affects narrow content.
- Callaway, H.G. 1992. Meaning holism and semantic realism. *Dialectica* 46:41-59.
- Churchland, P.M. 1998. Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal Of Philosophy* 95:5-32.

Devitt, M. 1994. A critique of the case for semantic holism. *Philosophical Perspectives* 8:281-306.

Devitt, M. 1994. Semantic localism: Who needs a principled basis? In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*. Holder-Pichler-Tempsky.

Esfeld, M. 1998. Holism and analytic philosophy. *Mind* 107:365-80.

Fodor, J.A. & LePore, E. 1992. *Holism: A Shopper's Guide*. Blackwell.
Rebutting arguments for meaning holism: those based on confirmation holism (Quine), normativity of interpretation (Davidson, Dennett, Lewis), and functional-role semantics (Block, Field, Churchland).

Fodor, J.A. & LePore, E. 1993. *Precis of Holism: A Shopper's Guide*. *Philosophy and Phenomenological Research* 53:637-682.
A discussion of *Holism* with comments by Devitt, Rey, McLaughlin, Brandom, and Churchland, and a reply by Fodor and Lepore.

Gauker, C. 1993. Holism without meaning: A critical review of Fodor and Lepore's *Holism: A Shopper's Guide*. *Philosophical Psychology* 6:441-49.

Harrell, M. 1996. Confirmation holism and semantic holism. *Synthese* 109:63-101.

Heal, J. 1994. Semantic holism: Still a good buy. *Proceedings of the Aristotelian Society* 68:325-39.
A critique of Fodor and Lepore. Distinguishes versions of holism, and argues for a weak version. Real thinkers are subjects, which imposes constraints on the interrelations of thoughts. Science fiction is irrelevant here.

Kukla, A. & Kukla, R. 1989. Meaning holism and intentional psychology. *Analysis* 173-53.
Contra Fodor, meaning holism is compatible with intentional psychology. Most psychological generalizations quantify over contents, rather than appealing to specific contents.

Lormand, E. 1996. How to be a meaning holist. *Journal of Philosophy* 93:51-73.

Margolis, E. & Laurence, S. 1998. Multiple meanings and stability of content. *Journal of Philosophy* 5:255-63.

McClamrock, R. 1989. Holism without tears: Local and global effects in cognitive processing. *Philosophy of Science* 56:258-74.

Miller, R.B. 1997. One bad and one not very good argument against holism. *Australasian Journal of Philosophy* 75:234-40.
A nice criticism of Fodor and Lepore's arguments that holism implies (a) the nonexistence of intentional laws and (b) the nonlearnability of language.

Pagin, P. 1997. Is compositionality compatible with holism? *Mind and Language* 12:11-33.

Perry, J. 1994. Fodor and Lepore on holism. *Philosophical Studies* 73:123-58.
The argument from anatomism and the failure of the analytic/synthetic distinction to holism fails. On the many different interpretations of holism and anatomism: there is a reasonable molecularist position.

Senor, T.D. 1992. Two-factor theories, meaning holism, and intentionalistic psychology: A reply to Fodor. *Philosophical Psychology* 5:133-51.

Silverberg, A. 1994. Meaning holism and intentional content. *Pacific Philosophical Quarterly* 75:29-53.

Stich, S.P. 1983. Some evidence against narrow causal theories of belief. In *From Folk Psychology to Cognitive Science*. MIT Press.

Talmage, C.J.L. & Mercer, M. 1991. Meaning holism and interpretability. *Philosophical Quarterly* 41:301-15.

Talmage, C.J.L. 1998. Semantic localism and the locality of content. *Erkenntnis* 48:101.

2.8 Mental Content, Misc

Allen, C. 1992. Mental content. *British Journal for the Philosophy of Science* 43:537-53.

Beckermann, A. 1996. Is there a problem about intentionality? *Erkenntnis* 45:1-24.

Bilgrami, A. 1992. *Belief and Meaning: The Unity and Locality of Mental Content*. Blackwell.

Bogdan, R.J. 1986. The manufacture of belief. In (R. Bogdan, ed) *Belief: Form, Content, and Function*. Oxford University Press.

Butler, K. 1995. Content, context, and compositionality. *Mind and Language* 10:3-24.

Cummins, R. 1991. Form, interpretation, and the uniqueness of content: A response to Morris. *Minds and Machines* 1:31-42.

Morris 1991 is wrong: formal individuation is easy, and objectively determinate content isn't needed. External grounding is also irrelevant.

Devitt, M. 1994. The methodology of naturalistic semantics. *Journal of Philosophy* 91:519-44.

Engel, P. 2000. Wherein lies the normative dimension in meaning and mental content? *Philosophical Studies* 100:305-321.

Fodor, J.A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Fodor, J.A. 1990. *A Theory of Content and Other Essays*. MIT Press.

Fodor, J.A. 1994. *The Elm and the Expert*. MIT Press.

Garfield, J.L. 2000. The meanings of "meaning" and "meaning": Dimensions of the sciences of mind. *Philosophical Psychology* 31:421-440.

Gillett, G. 1992. *Representation, Meaning, and Thought*. Oxford University Press.

Haldane, J.J. 1989. Naturalism and the problem of intentionality. *Inquiry* 32:305-22.

Haugeland, J. 1990. The Intentionality All-Stars. *Philosophical Perspectives* 4:383-427.

Intentionality around the diamond: neoCartesianism, neobehaviorism, neopragmatism. 1B=Fodor/Pylyshyn, 2B=Dennett/Quine, 3B=Heidegger/Sellars. SS=Wittgenstein. RF=Searle, CF=Skinner, LF=Rorty/Derrida. Lots of fun.

Horgan, T. 1994. Naturalism and intentionality. *Philosophical Studies* 76:301-26.

Jacob, P. 1997. *What Minds Can Do: Intentionality in a Non-intentional World*. Cambridge University Press.

Kaye, L.J. 1995. A scientific psychologicistic foundation for theories of meaning. *Minds and Machines* 5:187-206.

Lehrer, K. 1986. Metamind: Belief, consciousness and intentionality. In (R. Bogdan, ed) *Belief: Form, Content, and Function*. Oxford University Press.

Lycan, W.G. 1986. Thoughts about things. In (M. Brand & R. Harnish, ed) *The Representation of Knowledge and Belief*. University of Arizona Press.

Madell, G. 1989. Physicalism and the content of thought. *Inquiry* 32:107-21.

Maloney, J.C. 1989. *The Mundane Matter of the Mental Language*. Cambridge University Press.

Martin, C.B. & Pfeifer, K. 1986. Intentionality and the non-psychological. *Philosophy and Phenomenological Research* 46:531-54.

McDowell, J. 1998. Lecture III: Intentionality as a relation. *Journal Of Philosophy* 95:471-491.

McGinn, C. 1989. *Mental Content*. Blackwell.

McManus, D. 2000. Boghossian, Miller and Lewis on dispositional theories of meaning. *Mind and Language* 15:393-399.

Miller, A. 1997. Boghossian on reductive dispositionalism about content: The case strengthened. *Mind and Language* 12:1-10.

Morris, M. 1991. Why there are no mental representations. *Minds and Machines* 1:1-30.

There can be no non-stipulative content to non-semantically individuated tokens. Mostly a critique of Cummins; also Fodor and Dennett.

Newton, N. 1996. *Foundations of Understanding*. John Benjamins.

Pacherie, E. 2000. The content of intentions. *Mind and Language* 15:400-432.

Peacocke, C. 1986. *Thoughts: An Essay on Content*. Blackwell.

Peacocke, C. 1991. Content and norms in a natural world. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Pollock, J. 1990. Understanding the language of thought. *Philosophical Studies* 58:95-120.

Remarks on a number of aspects of mental content -- narrow, propositional, qualitative -- with respect to functionalism and the language of thought. With comments by Baker.

Priest, G. 2000. Objects of thought. *Australasian Journal of Philosophy* 78:494-502.

Prinz, J. 2000. The duality of content. *Philosophical Studies* 100:1-34.

Schiffer, S. 1981. Truth and the theory of content. In (H. Parret, ed) *Meaning and Understanding*. Berlin.

Schiffer, S. 1987. *Remnants of Meaning*. MIT Press.

Sellars, W. & Chisholm, R. 1957. Intentionality and the mental: A correspondence. *Minnesota Studies in the Philosophy of Science* 2:507-39.

Stalnaker, R. 1999. *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press.

Sterelny, K. 1990. *The Representational Theory of Mind*. Blackwell.

Stich, S.P. 1982. On the ascription of content. In (A. Woodfield, ed) *Thought and Object*. Oxford University Press.

On the tacit theories underlying the folk psychology of belief: beliefs are states associated with typical causal patterns. With remarks on ambiguity, content identity and similarity, and environmental dependence.

Stich, S.P., and Laurence, S. 1994. Intentionality and naturalism. *Midwest Studies in Philosophy* 19:159-82. Reprinted in (Stich) *Deconstructing the Mind*. Oxford University Press, 1996.

Argues that a failure to "naturalize" intentionality won't lead to disasters such as irrealism, irrelevance, or non-science, whether naturalization is understood as analysis, property identity, supervenience, or whatever.

Tye, M. 1994. Naturalism and the problem of intentionality. *Midwest Studies in Philosophy* 19:122-42.

There's no deep problem of naturalism about intentionality, as we know it's true already. The real puzzle is that of finding a mechanism to close the gap, e.g. via analysis or essentialism. But naturalism doesn't require that.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 3: Metaphysics of Mind [1026]

=====

- 3.1 Supervenience [94]
 - 3.1a Psychophysical Supervenience (Kim, etc) [14]
 - 3.1b Supervenience and Physicalism [25]
 - 3.1c Technical Issues in Supervenience [27]
 - 3.1d Supervenience, General [28]
- 3.2 Reduction [82]
 - 3.2a Multiple Realizability and Reduction [25]
 - 3.2b Nonreductive Materialism [28]
 - 3.2c Reduction in Psychology [14]
 - 3.2d Miscellaneous [15]
- 3.3 Other Psychophysical Relations [127]
 - 3.3a Physicalism [29]
 - 3.3b Token Identity [7]
 - 3.3c Emergence [58]
 - 3.3d Dualism [21]
 - 3.3e Miscellaneous [12]
- 3.4 Functionalism [63]
 - 3.4a Causal Role Functionalism (Armstrong/Lewis) [17]
 - 3.4b Machine Functionalism (Putnam) [15]
 - 3.4c Miscellaneous [31]
- 3.5 Other Psychophysical Theories [238]
 - 3.5a Logical Behaviorism (Ryle, etc) [53]
 - 3.5b Identity Theory (Smart, Place, etc) [98]
 - 3.5c Eliminative Materialism (Rorty, Feyerabend) [30]
 - 3.5d Anomalous Monism (Davidson) [57]
- 3.6 Mental Causation [58]
- 3.7 Personal Identity [189]
 - 3.7a Personal Identity, General [76]
 - 3.7b Parfit on Personal Identity [49]
 - 3.7c Persons [23]
 - 3.7d Split Brains [22]
 - 3.7e Multiple Personality Disorder [19]
- 3.8 Free Will [129]
- 3.9 The Problem of Other Minds [46]

3.1 Supervenience

3.1a Psychophysical Supervenience (Kim, etc)

Crane, T. 1991. Why indeed? Papineau on supervenience. *Analysis* 51:32-7.
Contra Papineau 1989: the assumption of completeness is false or trivial.
Maybe the mental is part of a complete physics. With response by Papineau.

Elugardo, R. 1988. Against weak psychophysical supervenience. *Dialectica* 42:129-43.
Various objections to Kim's arguments for supervenience. Not all internal states relevant to I/O relations are psychological states. Strange.

Kim, J. 1979. Causality, identity and supervenience in the mind-body problem. *Midwest Studies in Philosophy* 4:31-49.
Supervenience of the mental on the physical is what is required to make mental causation possible. Very nice.

Kim, J. 1982. Psychophysical supervenience. *Philosophical Studies* 41:51-70.
Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).
Internal mental states (i.e. ones that are not rooted outside) supervene on synchronous internal physical states, and internal states are all that is

relevant in the explanation of behavior.

Kim, J. 1982. Psychophysical supervenience as a mind-body theory. *Cognition and Brain Theory* 5:129-47.

Distinguishes weak (within-world) vs strong (across-worlds) supervenience.

Relates to reduction, internal/external mental states, and various theories.

Kim, J. 1997. Supervenience, emergence, and realization in the philosophy of mind. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Lewis, H. 1985. Is the mental supervenient on the physical? In (B. Vermazen & M. Hintikka, eds) *Essays on Davidson*. Oxford University Press.

On some problems with supervenience, the relation between supervenience and

reduction, and on reasons for accepting psychophysical supervenience. *Loose*.

Loar, B. 1993. Can we confirm supervenient properties? In (E. Villanueva, ed) *Naturalism and Normativity*. Ridgeview.

If mental properties are supervenient but irreducible to physical/functional properties, we can't confirm them. Confirmation requires an indispensable

explanatory role, which irreducibility precludes. With comments by Schiffer.

Macdonald, C. 1995. Psychophysical supervenience, dependency, and reduction.

In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Merricks, T. 1998. Against the doctrine of microphysical supervenience. *Mind* 107: 59-72.

Noonan, H. 1999. Microphysical supervenience and consciousness. *Mind*

108:755-9.

Papineau, D. 1989. Why supervenience? *Analysis* 50:66-71.

Psychophysical supervenience follows from completeness of physical laws.

Papineau, D. 1995. Arguments for supervenience and physical realization. In

(E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge

University Press.

Witmer, D.G. 1998. What is wrong with the manifestability argument for

supervenience? *Australasian Journal of Philosophy* 76:84-89.

3.1b Supervenience and Physicalism [see also 3.3a]

Armstrong, D.M. 1982. Metaphysics and supervenience. *Critica* 42:3-17.

Argues that everything is logically supervenient on the physical. Considers classes, possibilities, numbers, universals, and objects of thought.

Bailey, A. 1998. Supervenience and physicalism. *Synthese* 117:53-73.

Chalmers, D.J. 1996. Supervenience and materialism. In *The Conscious Mind*

(pp. 41-42). Oxford University Press, 1996.

Charles, D. 1992. Supervenience, composition, and physicalism. In (D. Charles

& K. Lennon, eds) *Reduction, Explanation and Realism*. Oxford University

Press.

Francescotti, R.M. 1998. Defining "physicalism". *Journal of Mind and Behavior*

19:51-64.

Haugeland, J. 1984. Ontological supervenience. *Southern Journal of Philosophy*

Supplement 22:1-12.

Supervenience is all we need for materialism. Various materialist arguments (unity, "nothing but", history, fear of darkness, simplicity, law) don't support physical exhaustion & token identity, over and above supervenience.

Hellman, G. & Thomson, F. 1975. Physicalism: ontology, determination and reduction. *Journal of Philosophy* 72:551-64.

Hellman, G. & Thomson, F. 1977. Physicalist materialism. *Nous* 11:309-45.
Some applications of the earlier treatment: examples of determination without reduction; the statuf of properties and universals; the mental; the life sciences; modalities and essentialism; theoretical equivalence.

Hellman, G. 1985. Determination and logical truth. *Journal of Philosophy* 82:607-16.
Some remarks on determination, physicalism, model theory, and logical truth.

Horgan, T. 1981. Token physicalism, supervenience, and the generality of physics. *Synthese* 49:395-413.
Argues that the generality of physics should be a supervenience thesis, not token physicalism. Fodor's token physicalism is untenable but might be saved with an appropriate view of events.

Horgan, T. 1982. Supervenience and microphysics. *Pacific Philosophical Quarterly* 63:29-43.
An account of how all facts supervene on microphysical facts, and how all intrinsic facts supervene on intrinsic microphysical facts.

Horgan, T. 1984. Supervenience and cosmic hermeneutics. *Southern Journal of Philosophy Supplement* 22:19-38.
Laplacean demon's job: number crunching, plus cosmic hermeneutics to explain high-level truths. All high-level truths follow from low-level by meaning constraints. Application to theoretical/mentalistic/everyday terms. Nice.

Jack, A. 1994. Materialism and supervenience. *Australasian Journal of Philosophy* 72:426-43.
Supervenience is neither necessary nor sufficient for materialism. With various (contentious) counterexamples. So we need a different formulation.

Kirk, R. 1996. Strict implication, supervenience, and physicalism. *Australasian Journal of Philosophy* 74:244-57.
Argues for strict implication rather than supervenience as a formulation of "minimal physicalism" (unless supervenience is formulated just right).

Lewis, D. 1983. New work for a theory of universals. *Australasian Journal of Philosophy*.
Formulates a definition of materialism: among worlds where no natural properties alien to our worlds are instantiated, no two differ without differing physically. With a lot of other material on universals.

Melnyk, A. 1991. Physicalism: From supervenience to elimination. *Philosophy and Phenomenological Research* 51:573-87.
How can supervenience, as a relationship between ontologically distinct properties, be explained? Modal realism and grand-properties don't work. Eliminativism about supervenient properties is the only possibility.

Moreland, J.P. 1999. Should a naturalist be a supervenient physicalist? *Metaphilosophy* 29:35-57.

Moser, P. & Trout, J.D. 1996. Physicalism, supervenience, and dependence. In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Pettit, P. 1993. A definition of physicalism. *Analysis* 53:213-23.

Defines physicalism in terms of claims that microphysical entities constitute everything and that microphysical laws govern everything. With a reply by Crane.

Rowlands, M. 1995. *Supervenience and Materialism*. Avebury.

Seager, W.E. 1988. Weak supervenience and materialism. *Philosophy and Phenomenological Research* 48:697-709.

Weak supervenience provides a more tenable form of materialism than strong supervenience, because of inverted spectrum possibilities, etc.

Snowdon, P.F. 1989. On formulating materialism and dualism. In (J. Heil, ed) *Cause, Mind, and Reality: Essays Honoring C. B. Martin*. Kluwer.

A construal of materialism in terms of constitution, not identity. Discusses the entailment between physical properties and mental properties; considers a nonreductive physicalism and a primitive dualism.

Sturgeon, S. 1998. Physicalism and overdetermination. *Mind* 107:411-32.

Wilson, J. 1999. How superduper does a physicalist supervenience need to be? *Philosophical Quarterly* 49:33-52.

Witmer, D.G. 1999. Supervenience physicalism and the problem of extras. *Southern Journal of Philosophy* 37:315-31.

3.1c Technical Issues in Supervenience

Bacon, J. 1986. Supervenience, necessary coextensions, and reducibility. *Philosophical Studies* 49:163-76.

A modal-logic analysis of the relations between various notions of supervenience. Most concepts of supervenience entail necessary co-extension, under certain closure assumptions for properties.

Bacon, J. 1995. Weak supervenience supervenes. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Bonevac, D. 1988. Supervenience and ontology. *American Philosophical Quarterly* 25:37-47.

A model-theoretic treatment of supervenience, in terms of relations between theories. Supervenience turns out to be equivalent to reduction.

Bovens, L. 1994. Principles of supervenience. *Australasian Journal of Philosophy* 72:294-301.

Divers, J. 1996. Supervenience for operators. *Synthese* 106:103-12.

Forrest, P. 1988. Supervenience: The grand-property hypothesis. *Australasian Journal of Philosophy* 66:1-12.

A nonreductive supervenience hypothesis: supervenient properties are properties of properties, e.g intrinsic goodness is a property of an object's nature.

Forrest, P. 1992. Universals and universalisability: An interpretation of Oddie's discussion of supervenience. *Australasian Journal of Philosophy* 70:93-98.

Grim, P. 1997. Worlds by supervenience: Some further problems. *Analysis* 2:146-51.

Grimes, T. 1991. Supervenience, determination, and dependency. *Philosophical Studies* 62:81-92.

On dependency supervenience (B properties determine A properties) versus

determination supervenience (A properties need B properties).

Grimes, T. 1995. The Tweedledum and Tweedledee of supervenience. In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Humberstone, I.L. 1992. Some structural and logical aspects of the notion of supervenience. *Logical Analysis* 35:101-37.

Kim, J. 1984. Concepts of supervenience. *Philosophy and Phenomenological Research* 45:153-76. Reprinted in *_Supervenience and Mind_* (Cambridge University Press, 1993).

Distinguishes weak and strong supervenience. A mistaken proof that strong and global supervenience are equivalent. Strong supervenience implies a kind of reduction, but not an explanatorily useful reduction.

Kim, J. 1987. 'Strong' and 'global' supervenience revisited. *Philosophy and Phenomenological Research* 48:315-26. Reprinted in *_Supervenience and Mind_* (Cambridge University Press, 1993).

Reasons why global supervenience doesn't entail strong supervenience, and trying to rescue global supervenience as a useful notion. Suggests a similarity-based notion of global supervenience.

Kim, J. 1988. Supervenience for multiple domains. *Philosophical Topics* 16:129-50. Reprinted in *_Supervenience and Mind_* (Cambridge University Press, 1993).

How properties in one domain can supervene on properties in another, with or without co-ordination between domains. Relation to global supervenience.

Klagge, J.C. 1995. Supervenience: Model theory or metaphysics? In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Marras, A. 1993. Supervenience and reducibility: An odd couple. *Philosophical Quarterly* 43:215-222.

Supervenience doesn't entail reducibility, as necessary coextension doesn't suffice, and is incompatible with reducibility, due to ontological asymmetry.

McLaughlin, B.P. 1995. Varieties of supervenience. In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Distinguishes modal-operator and possible-worlds versions of supervenience, and explicates global supervenience and its relation to weak and strong. With remarks on multiple-domain supervenience and the relation to reduction.

McLaughlin, B. P. 1997. Supervenience, vagueness, and determination. *Philosophical Perspectives* 11:209-30.

Melnyk, A. 1997. On the metaphysical utility of claims of global supervenience. *Philosophical Studies* 87:277-308.

Moser, P.K. 1992. Physicalism and global supervenience. *Southern Journal of Philosophy* 30:71-82.

Argues that global supervenience has epistemological problems -- how could we ever know that it holds, and that certain worlds are impossible?

Oddie, G. & Tichy, P. 1990. Resplicing properties in the supervenience base. *Philosophical Studies* 58:259-69.

Closure under resplicing makes supervenience both too narrow and too wide. Weak supervenience is generally too weak to capture the dependence relation.

Oddie, G. 1991. Supervenience and higher-order universals. *Australasian Journal of Philosophy* 69:20-47.

Paull, R.C. & Sider, T.R. 1992. In defense of global supervenience. *Philosophy and Phenomenological Research* 52:833-53.

Gives a proof of the distinction between strong and global supervenience that improves on Petrie's, and argues contra Kim that global supervenience is a perfectly reasonable dependence relation for physicalism.

Petrie, B. 1987. Global supervenience and reduction. *Philosophy and Phenomenological Research* 48:119-30.

Defending global supervenience: it's weaker than strong supervenience, as base properties of other individuals are relevant. It doesn't entail type or token reducibility. On the relation to implicit definability and reduction.

Post, J.F. 1995. "Global" supervenient determination: Too permissive? In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Shagrir, O. 1999. More on global supervenience. *Philosophy and Phenomenological Research* 59:691-701.

van Cleve, J. 1990. Supervenience and closure. *Philosophical Studies* 58:225-38.

Properties in supervenience relations shouldn't be closed under negation or resplicing, due to bad consequences. With reply by Bacon on resplicing.

3.1d Supervenience, General

Blackburn, S. 1984. Supervenience revisited. In (I. Hacking, ed) *_Exercises in Analysis: Essays by Students of Casimir Lewy_*. Cambridge University Press.

On the incompatibility of weak supervenience without strong supervenience and realism. With discussion of various strengths of necessity involved in supervenience claims, and application to moral realism and anomalous monism.

Currie, G. 1984. Individualism and global supervenience. *British Journal for the Philosophy of Science* 35:345-58.

How social facts supervene on the totality of individual facts. Application to belief, etc.

Enc, B. 1996. Nonreducible supervenient causation. In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Grimes, T. 1988. The myth of supervenience. *Pacific Philosophical Quarterly* 69:152-60.

Supervenience is too weak to function as a dependency relation, as e.g. it can hold in two directions at once.

Hare, R.M. 1984. Supervenience. *Proceedings of the Aristotelian Society* 58:1-16.

On the universal conditionals that underlie supervenience, and the necessity thereof. A discussion of the necessity of moral, natural kind, and other sorts of supervenience. Contra Davidson, anomalous supervenience is silly.

Heil, J. 1995. Supervenience redux. In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Hellman, G. 1992. Supervenience/determination a two-way street? Yes, but one of the ways is the wrong way! *Journal of Philosophy* 89:42-47.

Reply to Miller 1990. Miller underestimates the modal force of supervenience and invokes irrelevant dispositional properties.

Horgan, T. 1993. From supervenience to superdupervenience: Meeting the demands of a material world. *Mind* 102:555-86.

An overview of supervenience, with focus on the problem of explaining

supervenience relations. With remarks on mental causation, emergence, physicalism, and reduction.

Kim, J. 1978. Supervenience and nomological incommensurables. *American Philosophical Quarterly* 15:149-56.

Developing and motivating the notion of supervenience. Investigating the relationship to reducibility and definability (equivalence, under certain conditions), and to microphysical determination.

Kim, J. 1984. Supervenience and supervenient causation. *Southern Journal of Philosophy Supplement* 22:45-56.

On weak/strong supervenience, and high-level causation via supervenience.

Kim, J. 1991. Supervenience as a philosophical concept. *Metaphilosophy* 21:1-27. Reprinted in *_Supervenience and Mind_* (Cambridge University Press, 1993).

A nice overview of supervenience and covariance.

Kim, J. 1993. *_Supervenience and Mind_*. Cambridge University Press.

A collection of articles on supervenience and causation in metaphysics and the philosophy of mind, with some added postscripts.

Kincaid, H. 1987. Supervenience doesn't entail reducibility. *Southern Journal of Philosophy* 25:343-56.

Supervenience doesn't entail reducibility, which is epistemological. The problem's not just huge disjuncts, but also the sharing of bases, no local correlations, and base-properties presupposing supervenient properties.

Kincaid, H. 1988. Supervenience and explanation. *Synthese* 77:251-81.

Argues that lower-level theories can explain supervenient but irreducible higher-level theories, but only under certain conditions, as low-level accounts don't have the relevant kind terms.

Klagge, J.C. 1988. Supervenience: Ontological and ascriptive. *Australasian Journal of Philosophy* 66:461-70.

On supervenience as an ontological relation (via metaphysical necessity) or as an ascriptive relation (via conceptual necessity). The first doesn't preclude the second. Moral realism and mental realism are in the same boat.

Loewer, B. 1995. An argument for strong supervenience. In (E. Savellos & U. Yalchin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

McLaughlin, B.P. 1983. Event supervenience and supervenient causation. *Southern Journal of Philosophy Supplement* 22:71-91.

McLaughlin, B.P. 1994. Varieties of supervenience. In (E. Savellos & O. Yalchin, eds) *_Supervenience: New Essays_*.

On a number of issues: possible worlds vs modal notions, explicating global supervenience, the relation between weak/strong/global supervenience, multiple-domain supervenience, and implications for reduction.

Melnyk, A. 1997. On the metaphysical utility of claims of global supervenience. *Philosophical Studies* 87:277-308.

Merricks, T. 1998. Against the doctrine of microphysical supervenience. *Mind* 107:59-71.

Miller, R.B. 1990. Supervenience is a two-way street. *Journal of Philosophy* 87:695-701.

If supervening properties can make arbitrarily fine distinctions, then physical properties supervene on moral/aesthetic/mental properties.

Noonan, H. 1987. Supervenience. *Philosophical Quarterly* 37:78-85.

Contra Blackburn 1984 on the possibility of weak supervenience without strong supervenience, even with metaphysical necessity; using Nozick's concept structures, or indexical definitions. With application to moral realism.

Post, J.F. 1984. On the determinacy of valuation. *Philosophical Studies* 45:315-33.

Stalnaker, R. 1996. Varieties of supervenience. *Philosophical Perspectives* 10:221-42.

Distinguishes "reductionist" and "metaphysical" conceptions of supervenience. Also discusses the relation between strong and global supervenience, degrees of necessity, and the explanatory role of supervenience.

Teller, P. 1984. The poor man's guide to supervenience and determination. *Southern Journal of Philosophy Supplement* 22:137-62.

Compares the Hellman/Thompson notion of determination with Kim's development of supervenience. Uses these to investigate the concept of materialism, and argues that materialism isn't contingent.

Teller, P. 1985. Is supervenience just disguised reduction? *Southern Journal of Philosophy* 23:93-100.

van Brakel, J. 1996. Interdiscourse or supervenience relations: The primacy of the manifest image. *Synthese* 106:253-97.

Zangwill, N. 1997. Explaining supervenience: Moral and mental. *Journal of Philosophical Research* 22:509-18.

3.2 Reduction

3.2a Reduction and Multiple Realizability

Antony, L.M. & Levine, J. 1997. Reduction with autonomy. *Philosophical Perspectives* 11:83-105.

<
Bechtel, W., & Mundale, J. 1999. Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science* 66:175-207.

Bickle, J. 1992. Multiple realizability and psychophysical reduction. *Behavior and Philosophy* 20:47-58.

Bickle, J. 1995. Connectionism, reduction, and multiple realizability. *Behavior and Philosophy* 23:29-39.

Block, N. 1997. Anti-reductionism slaps back. *Philosophical Perspectives* 11:107-32.

Bolender, J. 1995. Is multiple realizability compatible with antireductionism? *Southern Journal of Philosophy* 33:129-42.

Endicott, R.P. 1991. Macdonald on type reduction via disjunction. *Southern Journal of Philosophy* 29:209-14.

Endicott, R.P. 1989. On physical multiple realization. *Pacific Philosophical Quarterly* 70:212-24.

Endicott, R.P. 1993. Species-specific properties and more narrow reductive strategies. *Erkenntnis* 38:303-21.

On species-specific reductions. These can't reduce standard psychological properties, and problems with intra-species multiple realization can't be circumvented without giving up property reduction for token event identity.

Endicott, R.P. 1998. Collapse of the new wave. *Journal of Philosophy* 95:53-72.

Fodor, J. 1997. Special sciences: Still autonomous after all these years. *Philosophical Perspectives* 11:149-63.

Francescotti, R.M. 1997. What multiple realizability does not show. *Journal of Mind and Behavior* 18:13-28.

The anti-reductionist argument assumes that functional properties aren't physical properties (not even extrinsic physical properties). This, not multiple realizability, does the work.

Heil, J. 1999. Multiple realizability. *American Philosophical Quarterly* 36:189-208.

Kim, J. 1992. Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research* 52:1-26. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Multiple realization is compatible with reductionism. Jade (= jadeite or nephrite) isn't a scientific kind, and neither are multiply realizable mental properties. So there's no global psychology, just lots of local reductions.

Kistler, M. 1999. Multiple realization, reduction and mental properties. *International Studies in the Philosophy of Science* 13.

Nelson, A. 1985. Physical properties. *Pacific Philosophical Quarterly* 66:268-82.

Some comments on Wilson 1985: some special-science properties may be relevantly different in kind from his expanded physical properties.

Macdonald, C. 1992. Psychological type-type reduction via disjunction. *Southern Journal of Philosophy* 30:65-69.

Mucciolo, L. 1974. The identity thesis and neuropsychology. *Nous* 8:327-42.

Argues contra Fodor and Block that neurological equipotentiality doesn't refute type materialism. Mental states may not be anatomically defined neural states, but they may be more abstract neural holograms.

Nasrin, M. 2000. Multiple realizability: Also a difficulty for functionalism. *Journal of Consciousness Studies* 7:25-34.

Ross, P.A. 1999. The limits of physicalism. *Philosophy of Science* 66:94-116.

Schwartz, J. 1992. Who's afraid of multiple realizability?: Functionalism, reductionism, and connectionism. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

Shagrir, O. 1998. Multiple realization, computation and the taxonomy of psychological states. *Synthese* 114:445-461.

Shapiro, L. 2000. Multiple realizations. *The Journal of Philosophy* 97:635-654.

Wilson, M. 1985. What is this thing called 'pain'? -- The philosophy of science behind the contemporary debate. *Pacific Philosophical Quarterly* 66:227-67.

Argues for type-type identities and for an expanded view of the physical, as properties from physics exhibit the same sort of multiple realizability as functional properties. Sophisticated, with many interesting examples.

Zangwill, N. 1995. Supervenience, reduction, and infinite disjunction. *Philosophia* 24:321-30.

3.2b Nonreductive Materialism [see also 3.5d]

Barrett, J. 1995. Causal relevance and nonreductive physicalism. *Erkenntnis* 42:339-62.

Beckermann, A. 1992. Reductive and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.

On varieties of physicalism with respect to reduction: semantic physicalism, identity theory, supervenience, and the denial of emergence. Advocates a version on which physical states realize mental states.

Beckermann, A, Flohr, H. & Kim, J. (eds) 1992. *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.

Boyd, R. 1980. Materialism without reductionism: What physicalism does not entail. In (N. Block, ed) *_Readings in the Philosophy of Psychology_*, Vol 1. MIT Press.

Dupre, J. 1988. Materialism, physicalism, and scientism. *Philosophical Topics* 16:31-56.

Arguing for a pluralistic conception. With criticism of Churchland's reductionism, Davidson's token identity, and more generally reverential "scientism". Reductionist explanation is not the general rule.

Ellis, R. 2000. Consciousness, self-organization, and the process-substratum relation: Rethinking nonreductive physicalism. *Philosophical Psychology* 13:173-190.

Fodor, J.A. 1974. Special sciences. *Synthese* 28:97-115. Reprinted in (N. Block, ed) *_Readings in the Philosophy of Psychology_* (MIT Press, 1980).

Psychological kinds can't be reduced to physical kinds, due to cross-classification, although token physicalism still holds. How to maintain the generality of physics without a reductionist unity of science.

Francescotti, R.M. 1998. The nonreductionist's troubles with supervenience. *Philosophical Studies* 89:105-24.

Horgan, T. 1993. Nonreductive materialism and the explanatory autonomy of psychology. In (S. Wagner & R. Warner, eds) *_Naturalism: A Critical Appraisal_*. University of Notre Dame Press.

Gives four constraints on interlevel connections, and some arguments against reductionism and for the autonomy of psychology. Argues that supervenience fact are themselves in need of explanation.

Horgan, T. 1994. Nonreductive materialism. In (R. Warner & T. Szubka, eds) *_The Mind-Body Problem: A Guide to the Current Debate_*. Blackwell.

Kernohan, A. 1988. Non-reductive materialism and the spectrum of mind-body identity theories. *Dialogue* 27:475-88.

Classifying psychophysical theories by the status (necessary, lawful, anomalous, false) of psychophysical/psychological generalizations. Defending autonomous monism: nonreductive materialism with psychological laws.

Kim, J. 1989. The myth of non-reductive materialism. *Proceedings and Addresses of the American Philosophical Association* 63(3):31-47. Reprinted in *_Supervenience and Mind_* (Cambridge University Press, 1993).

Somewhat loose arguments that non-reductive physicalist realism is untenable. Anomalous monism makes the mental irrelevant, functionalism is compatible with species-specific reduction, and supervenience is weak or reductive.

Kim, J. 1992. The nonreductivist's trouble with mental causation. In (J. Heil & A. Mele, eds) *_Mental Causation_*. Oxford University Press. Reprinted in

Supervenience and Mind (Cambridge University Press, 1993).

Argues that nonreductive materialism implies downward causation (as the mental has more causal powers than the physical alone), and that downward causation violates the causal closure of the physical.

Kim, J. 1992. "Downward causation" in emergentism and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) Emergence or Reduction?: Prospects for Nonreductive Physicalism. De Gruyter.

Argues that nonreductive materialism is just like 1930s emergentism, with the the mental contributing new causal powers, and so implies downward causation.

Kirk, R. 1996. How physicalists can avoid reductionism. *Synthese* 108:157-70.
Contra Kim, physicalists can avoid reduction by embracing strict implication.

Margolis, J. 1978. Persons and Minds: The Prospects of Non-Reductive Materialism. D. Reidel.

Marras, A. 1993. Psychophysical supervenience and nonreductive materialism. *Synthese* 95:275-304.

Marras, A. 1994. Nonreductive materialism and mental causation. *Canadian Journal of Philosophy* 24:465-93.

Melnyk, A. 1995. Two cheers for reductionism, or, the dim prospects for nonreductive materialism. *Philosophy of Science* 62:370-88.

Melnyk, A. 1998. The prospects for Kirk's nonreductive physicalism. *Australasian Journal of Philosophy* 76:323-32.

Loar, B. 1992. Elimination versus nonreductive physicalism. In (D. Charles & K. Lennon, eds) Reduction, Explanation and Realism. Oxford University Press.

Papineau, D. 1992. Irreducibility and teleology. In (D. Charles & K. Lennon, eds) Reduction, Explanation and Realism. Oxford University Press.
Non-reductive physicalism is a mystery unless we invoke teleology.

Pereboom, D. & Kornblith, H. 1991. The metaphysics of irreducibility. *Philosophical Studies* 63:125-45.

Explicating anti-reductionism: mental causal powers are constituted of physical causal powers, but aren't type- or token-identical to them. Against arguments from local reduction, neuroscience, explanatory exclusion, etc.

Silvers, S. 1997. Nonreductive naturalism. *Theoria* 12:163-84.

Smith, A.D. 1993. Non-reductive physicalism? In (H. Robinson, ed) Objections to Physicalism. Oxford University Press.

A careful discussion of how to characterize physicalism, in terms of identity or supervenience, and argues that physicalism must reduce (bowdlerize) qualia to something they are not, as physicalism requires topic-neutral analyses.

Ten Elshof, G. 1997. Supervenient difficulties with nonreductive physicalism: A critical analysis of supervenience physicalism. *Kinesis* 24:3-22.

van Gulick, R. 1992. Nonreductive materialism and the nature of intertheoretical constraint. In (A. Beckermann, H. Flohr, & J. Kim, eds) Emergence or Reduction?: Prospects for Nonreductive Physicalism. De Gruyter.
On how a nonreductive materialism can handle problems about mental causation, psychophysical dependencies, and qualia. A teleofunctionalist view with different conceptual frameworks, but mental properties physically realized.

Wedgwood, R. 2000. The price of non-reductive physicalism. *Nous* 34:400-421.

3.2c Reduction in Psychology

Bickle, J. 1995. Psychoneural reduction of the genuinely cognitive: Some accomplished facts. *Philosophical Psychology* 8:265-85.
Argues that cognitive theories have already been reduced to neurobiology in some domains, such as associative learning.

Churchland, P.M. 1982. Is 'thinker' a natural kind? *Dialogue* 21:223-38.
Psychology shouldn't be autonomous from natural science. By analogy with biology, nature provides (a) conceptual insight, and (b) real constraints, e.g. thermodynamic ones. Biology and psychology are continuous.

Gaito, J. 1960. Description, explanation, and reductionism in psychology. *Psychological Reports* 6:203-5.

Gaito, J. & Leonard, D. 1965. Philosophical and empirical reductionism in psychology. *Journal of General Psychology* 72:69-75.

Hardcastle, V.G. 1992. Reduction, explanatory extension, and the mind/brain sciences. *Philosophy of Science* 59:408-28.
The relationship between psychology and neuroscience is best characterized not by reduction but by explanatory extension, where each field is enriched by the other. With a number of examples from recent empirical work.

Hyland, M.E. 1995. Against nomological reductionism in psychology: A response to Robinson. *New Ideas in Psychology* 13:9-11.

Jessor, R. 1958. The problem of reductionism in psychology. *Psychological Review* 65:170-78.

Marras, A. 1990. Reduction in psychology. *Acta Analytica* 6:65-78.

Martindale, R.L. & Seidel, R.J. 1959. Reductionism: Its prodigal encores. *Psychological Reports* 5:213-16.

Montgomery, R. 1990. The reductionist ideal in cognitive psychology. *Synthese* 85:279-314.
Anti-reductionism needn't be ad hoc (contra Churchland). Although evolution provides some pressure for 1-1 psychophysical mappings, there are significant countervailing forces, e.g. in vision, memory, learning, and language use.

Olszewsky, T.M. 1975. Dispositions and reductionism in psychology. *Journal for the Theory of Social Behavior* 5:129-44.

Putnam, H. 1974. Reductionism and the nature of psychology. *Cognition* 2:131-46.

Richardson, R.C. 1999. Cognitive science and neuroscience: New wave reductionism. *Philosophical Psychology* 12:297-307.

Sloane, E.H. 1945. Reductionism. *Psychological Review* 52:214-23.

3.2d Reduction, Misc

Beckermann, A. 1997. Property physicalism, reduction, and realization. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Bickle, J. 1996. New wave psychophysical reductionism and the methodological caveats. *Philosophy and Phenomenological Research* 56:57-78.

Bickle, J. 1997. *Psychoneural Reductionism: The New Wave*. MIT Press.

Brooks, D.H.M. 1994. How to perform a reduction. *Philosophy and Phenomenological Research* 54:803-14.

Reduction comes to supervenience plus explicability. Thus biconditionals, multiple realizability, etc, are irrelevant. Biology is already reduced (mostly via functional explanation), and psychology looks promising. Nice.

Bunzl, M. 1987. Reductionism and the mental. *American Philosophical Quarterly* 24:181-9.

On the links between supervenience, reduction, and explanation. Supervenience is compatible with reductive explanation of a localized variety. We don't need laws, but explanatory links.

Causey, R.L. 1972. Attribute identities in microreductions. *Journal of Philosophy* 69:407-22.

Combes, R. 1988. Ockhamite reductionism. *International Philosophical Quarterly* 28:325-36.

Foss, J. 1995. Materialism, reduction, replacement, and the place of consciousness in science. *Journal of Philosophy* 92:401-29.

Hill, C.S. 1984. In defense of type materialism. *Synthese* 59:295-320.

Kitcher, P.S. 1980. How to reduce a functional psychology. *Philosophy of Science* 47:134-40.

Contra Richardson 1979, a purely functional psychology is irreducible. The genetics analogy is misleading; multiple realizations can't explain high-level laws.

Papineau, D. 1985. Social facts and psychological facts. In (G. Currie & A. Musgrave, eds) *_Popper and the Human Sciences_*. Martinus Nijhoff.

Mind is not reducible to body, but societies reduce to individuals. Multiple realization is in tension with predictability. Natural selection resolves the tension for the mental, but cannot for the social.

Richardson, R.C. 1979. Functionalism and reductionism. *Philosophy of Science* 46:533-58.

Argues that functionalism is compatible with reductionism, by analogies. Genetics has multiple realization and multiple function; reduction doesn't require biconditionals. With remarks on the de facto autonomy of psychology.

Richardson, R.C. 1982. How not to reduce a functional psychology. *Philosophy of Science* 49:125-37.

Response to Kitcher 1980. Reductions are usually domain-specific, and high-level regularities are indeed explained.

Sarkar, S. 1992. Models of reduction and categories of reductionism. *Synthese* 91:167-94.

Wimsatt, W. 1976. Reductionism, levels of organization, and the mind-body problem. In (G. Globus, ed) *_Consciousness and the Brain_*. Plenum Press.

Excellent coverage of the notion of level and its applicability to mind.

3.3 Other Psychophysical Relations

3.3a Physicalism [see also 1.3, 1.7b, 3.1b, 3.2b, 3.5]

Crane, T. 1991. All God has to do. *Analysis* 51:235-44.

If there are no contingent psychophysical laws, then there are no mental properties. So physicalism/supervenience is false; God had extra work to do.

Crane, T. 1993. A definition of physicalism: Reply to Pettit. *Analysis* 53:224-27.

Crane, T. & Mellor, D.H. 1990. There is no question of physicalism. *Mind* 99:185-206.

Physical sciences have no ontological authority over the mental. Considers and dismisses arguments from laws, causation, reduction, supervenience.

Daly, C. 1995. Does physicalism need fixing? *Analysis* 55:135-41.

Francescotti, R. 2000. Ontological physicalism and property pluralism: Why they are incompatible. *Pacific Philosophical Quarterly* 81:349-362.

Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*. Holder-Pichler-Tempsky.

On why materialism requires conceptual analysis to locate mental properties in the natural world. Even a posteriori necessary connections have to be backed by a priori links. With remarks on supervenience. A nice paper.

Kirk, R. 1979. From physical explicability to full-blooded materialism. *Philosophical Quarterly* 29:229-37.

If every physical events has a physical explanation, and the mental is causally efficacious, then mental facts are strictly implied by physical facts. A nice argument.

Kirk, R. 1982. Physicalism, identity, and strict implication. *Ratio* 24:131-41.

Materialism doesn't need a identity thesis. The requirement that mental facts are entailed by physical facts plays the role played by Kripke's requirement of necessary identity, and is more reasonable.

Kirk, R. 1996. Physicalism lives. *Ratio* 9:85-89.

Nothing in the arguments of Crane and Mellor 1990 count against a physicalism based on strict implication.

Madell, G. 1988. *Mind and Materialism*. Edinburgh University Press.

On the problems posed for materialism by intentionality, autonomy, awareness, and indexicality. Tentatively advocates a Cartesian position.

McGinn, C. 1980. Philosophical materialism. *Synthese* 44:173-206. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Melnyk, A. 1994. Being a physicalist: How and (more importantly) why.

Advocates "realization physicalism": all properties are either physical or functional properties realized by physical ones. This achieves unity between sciences better than alternatives, and avoids overdetermination.

Melnyk, A. 1996. Formulating physicalism: Two suggestions. *Synthese* 105:381-407.

Discusses two formulations of physicalism: requiring high-level properties to be disjunctions of physical states, or to be functional properties realized physically. Tentatively endorses the latter.

Melnyk, A. 1997. How to keep the 'physical' in physicalism. *Journal of Philosophy* 94:622-637.

Montero, B. 1999. The body problem. *Nous* 33:183-200.

Moser, P.K. 1996. Physicalism and mental causes: Contra Papineau. *Analysis* 56:263-67.

Nagel, E. 1949. Are naturalists materialists? *Journal of Philosophy*

42:515-53.

Papineau, D. 1994. Philosophical Naturalism. Blackwell.

Pettit, P. 1993. A definition of physicalism. *Analysis* 53:213-23.
Physicalism is the claim that (1) There are microphysical entities, (2)
Microphysical entities constitute everything, (3) There are microphysical
regularities, (4) Microphysical regularities govern everything.

Pettit, P. 1994. Microphysicalism without contingent micro-macro laws.
Analysis 54:253-57.

Pettit, P. 1995. Microphysicalism, dottiism, and reduction. *Analysis*
55:141-46.

Poland, J. 1994. Physicalism: The Empirical Foundations. Oxford University
Press.

Ravenscroft, I. 1997. Physical properties. *Southern Journal Of
Philosophy* 35:419-431.

Robinson, D. 1991. On Crane and Mellor's argument against physicalism. *Mind*
100:135-36.

Robinson, H. (ed) 1993. Objections to Physicalism. Oxford University Press.

Sheldon, W.H. 1946. Are naturalists materialists? *Journal of Philosophy*
43:197-209.

Sober, E. 1999. Physicalism from a probabilistic point of view. *Philosophical
Studies* 95:135-74.

Sturgeon, S. 1998. Physicalism and overdetermination. *Mind* 107:411-432.

Wilkes, K.V. 1973. Physicalism. Routledge and Kegan Paul.

3.3b Token Identity [see also 3.5b, 3.5d]

Foster, J. 1994. The token-identity thesis. In (R. Warner & T. Szubka, eds)
The Mind-Body Problem: A Guide to the Current Debate. Blackwell.

Horgan, T. & Tye, M. 1985. Against the token identity theory. In
(B. McLaughlin & E. LePore, eds) Action and Events. Blackwell.
We individuate mental events by their causal role, but we can't individuate
causes uniquely. So each mental event has multiple physical correlates, and
token identity doesn't hold.

Hornsby, J. 1981. Which physical events are mental events? *Proceedings of
the Aristotelian Society* 55:73-92.

Haugeland, J. 1982. Weak supervenience. *American Philosophical Quarterly*
19:93-103.
Supervenience doesn't imply token identity, and Davidson's argument for token
identity equivocates on "event". But weak supervenience (mentally discernible
worlds are physically discernible) is all we need. With nice examples.

Leder, D. 1985. Troubles with token identity. *Philosophical Studies* 47:79-94.
Physical/psychological token identity is no good: you can't individuate
physical events without psychological predicates.

Lurie, Y. 1978. Correlating brain states with psychological phenomena.
Australasian Journal of Philosophy 56:135-44.
Can't isolate the physical token of a belief, say, as it's always accompanied

by other beliefs. Meaning doesn't come in discrete tokens.

Peacocke, C. 1979. Argument for token identity. In *_Holistic Explanation_*. Oxford University Press.

3.3c Emergence

Alexander, S. 1920. *_Space, Time, and Deity_*. Macmillan.

Atkin A. 1992. On consciousness: What is the role of emergence? *Medical Hypotheses* 38:311-14.

Beckermann, A. 1992. Supervenience, emergence, and reduction. In (A. Beckermann, H. Flohr, & J. Kim, eds) *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.

On varieties of supervenience and of emergence, and of what is required for reduction. Argues that reduction involves general explanatory connections, whereas emergence involves unique and ultimate bridge laws.

Beckermann, A, Flohr, H. & Kim, J. (eds) 1992. *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.

Bedau, M. 1997. Weak emergence. *Philosophical Perspectives* 11:375-399.

Berenda, C.W. 1953. On emergence and prediction. *Journal of Philosophy* 50:269-74.

Bergmann, G. 1944. Holism, historicism, and emergence. *Philosophy of Science* 11:209-21.

Broad, C.D. 1925. *_The Mind and its Place in Nature_*. Routledge and Kegan Paul.

Bruntrup, G. 1998. Is psychophysical emergentism committed to dualism? The causal efficacy of emergent mental properties. *Erkenntnis* 48:133-51.

Bunge, M. 1977. Emergence and the mind. *Neuroscience* 2:501-9.

Garnett, A.C. 1942. Scientific method and the concept of emergence. *Journal of Philosophy* 39:477-86.

Haldane, J. 1996. The mystery of emergence. *Proceedings of the Aristotelian Society* 96:261-67.

A defence of radical emergence against Spencer-Smith 1995.

Hasker, W. 1982. Emergentism. *Religious Studies* 18:473-88.

Hasker, W. 1999. *_The Emergent Self_*. Cornell University Press.

Henle, P. 1942. The status of emergence. *Journal of Philosophy* 39:486-93.

Humphreys, P. 1996. Aspects of emergence. *Philosophical Topics* 24:53-71.

Humphreys, P. 1997. How properties emerge. *Philosophy of Science* 64:1-17.

Humphreys, P. 1997. Emergence, not supervenience. *Philosophy of Science Supplement* 64:337-45.

Jones, D.H. 1972. Emergent properties, persons, and the mind-body problem. *Southern Journal of Philosophy* 10:423-33.

Kekes, J. 1966. Physicalism, the identity theory, and the concept of emergence. *Philosophy of Science* 33:360-75.

- Kim, J. 1999. Making sense of emergence. *Philosophical Studies* 95:3-36.
- Klee, J. 1984. Microdeterminism and concepts of emergence. *Philosophy of Science* 51:44-63.
- Lovejoy, A.O. 1927. The meanings of "emergence" and its modes. In (E.S. Brightman, ed) *Proceedings of the Sixth International Congress of Philosophy*. Longmans, Green, and Co.
- Lowe, E. J. 2000. Causal closure principles and emergentism. *Philosophy* 75:571-586.
- Lowry, A. 1974. A note on emergence. *Mind* 83:276-77.
- Mackenzie, W.L. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:56-68.
- Margolis, J. 1986. Emergence. *Philosophical Forum* 17:271-95.
- McLaughlin, B.P. 1992. The rise and fall of British emergentism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.
A careful account of British emergentism. Explicates their view of emergent causal powers and laws in terms of fundamental configurational forces, a coherent idea that turned out to be false. An excellent paper.
- Meehl, P.E. & Sellars, W. 1956. The concept of emergence. In (H. Feigl & M. Scriven, eds) *Minnesota Studies in the Philosophy of Science*, vol. 1. University of Minnesota Press.
- Morgan, C.L. 1923. *Emergent Evolution*. Williams and Norgate.
- Morris, C.R. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:49-55.
- Newman, D. 1996. Emergence and strange attractors. *Philosophy of Science* 63:245-61.
- Newman, D.V. 2001. Chaos, emergence, and the mind-body problem. *Australasian Journal of Philosophy* 79:180-96.
- O'Connor, T. 1994. Emergent properties. *American Philosophical Quarterly* 31:91-104.
Argues against Alexander's and van Cleve's accounts of emergence, instead suggesting an account in terms of supervenience, non-structurality, and downward causation.
- Pap, A. 1951. The concept of absolute emergence. *British Journal for the Philosophy of Science* 2:302-11.
- Pepper, S.C. 1926. Emergence. *Journal of Philosophy* 23:241-45.
- Peters, S.L. 1995. *Emergent Materialism: A Proposed Solution to the Mind-Body Problem*. University Press of America.
- Pihlstrom, S. 1999. What shall we do with emergence? A survey of a fundamenta; issue in the metaphysics and epistemology of science. *South African Journal of Philosophy* 18:192-210.
- Pluhar, E. 1978. Emergence and reduction. *Studies in History and Philosophy of Science* 9:279-89.
- Ripley, C. 1984. Sperry's concept of consciousness. *Inquiry* 27:399-423.

An in-depth analysis of Sperry's views on consciousness. Sperry is not a dualist; he believes in "structural causation" based on emergent properties.

Rohrlich, F. 1997. Cognitive emergence. *Philosophy of Science Supplement* 64:346-58.

Rueger, A. 2000. Robust supervenience and emergence. *Philosophy of Science* 67:466-491.

Russell, E.S. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:39-48.

Schroder, J. 1998. Emergence: Non-deducibility or downwards causation? *Philosophical Quarterly* 48:433-52.

Silberstein, M. 1998. Emergence and the mind-body problem. *Journal of Consciousness Studies* 5:464-82.

Silberstein, M. & McGeever, J. 1999. The search for ontological emergence. *Philosophical Quarterly* 49:182-200.

Smart, J.J.C. 1981. Physicalism and emergence. *Neuroscience* 6:109-13.

Spencer-Smith, R. 1995. Reductionism and emergent properties. *Proceedings of the Aristotelian Society* 95:113-29.

Distinguishes radical, epistemic, and interactional emergence, favoring the latter. With consideration of qualia as a radical emergent.

Sperry, R.W. 1969. A modified concept of consciousness. *Psychological Review* 76:532-36.

Consciousness is an emergent property of brain dynamics that itself governs low-level flow of excitation. Midway between mentalism and materialism.

Sperry, R.W. 1991. In defense of mentalism and emergent interaction. *Journal of Mind and Behavior* 12:221-245.

Stace, W.T. 1939. Novelty, indeterminism, and emergence. *Philosophical Review* 48:296-310.

Stephan, A. 1992. Emergence -- a systematic look at its historical facets. In (A. Beckermann, H. Flohr, & J. Kim, eds) *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.

On different ways of understanding emergence: as nonadditivity, novelty, nonpredictability, nondeducibility; and on problems about qualia and downward causation.

Stephan, A. 1997. Armchair arguments against emergence. *Erkenntnis* 46:305-14.

Teller, P. 1992. A contemporary look at emergence. In (A. Beckermann, H. Flohr, & J. Kim, eds) *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.

An attempt to explicate "emergent" properties in terms of relational properties. Argues that even problem cases, e.g. space-time separation and phenomenal properties, might be treated this way.

van Cleve, J. 1990. Mind -- dust or magic? Panpsychism versus emergence. *Philosophical Perspectives* 4:215-226.

On Nagel 1979: emergence is more plausible than panpsychism. A construal of emergence as nomological supervenience without logical supervenience.

vandervert, L.R. 1991. On the modeling of emergent interaction: Which will it be, the laws of thermodynamics or Sperry's "wheel" in the subcircuitry?

Journal of Mind and Behavior 12:535-39.

Wimsatt, W.C. 1997. Aggregativity: Reductive heuristics for finding emergence. *Philosophy of Science* 64:372-84.

Wynn, M. 1999. Emergent phenomena and theistic explanation. *International Philosophical Quarterly* 39:141-55.

3.3d Dualism [see also 1.3f, 1.4e, 1.4f]

Averill, E.W. & Keating, B. 1981. Does interactionism violate a law of classical physics? *Mind* 90:102-7.

Interactionism is compatible with conservation of energy and momentum: the mind exerts a non-physical force on the brain.

Bricke, J. 1975. Interaction and physiology. *Mind* 84:255-9.

Efron, A. 1992. Residual asymmetric dualism: A theory of mind-body relations. *Journal of Mind and Behavior* 13:113-36.

Evans, S. 1981. Separable souls: A defense of minimal dualism. *Southern Journal of Philosophy* 19.

Herbert, R.T. 1998. Dualism/materialism. *Philosophical Quarterly* 48:159-75.

Larmer, R. 1986. Mind-body interactionism and the conservation of energy. *International Philosophical Quarterly* 26:277-85.

Various arguments about interactionism based on conservation of energy. C of E only applies to causally isolated systems, so objections beg the question.

Lowe, E.J. 1992. The problem of psychophysical causation. *Australasian Journal of Philosophy* 70:263-76.

Argues that there can be interaction without breaking physical laws: e.g. by basic psychic forces, or by varying physical constants, or especially by arranging fractal trees of physical causation leading to behavior.

Lowe, E.J. 1993. The causal autonomy of the mental. *Mind* 102:629-44.

Mills, E. 1996. Interactionism and overdetermination. *American Philosophical Quarterly* 33:105-115.

Argues that interactionist dualism is compatible with the causal closure of the physical, if we allow causal overdetermination; and there is a strong case for the latter.

Mills, E. 1997. Interactionism and physicality. *Ratio* 10:169-83.

O'Leary-Hawthorne, J. & McDonough, J.K. 1998. Numbers, minds, and bodies: A fresh look at mind-body dualism. *Philosophical Perspectives* 12:349-371.

Pap, A. 1952. Semantic analysis and psychophysical dualism. *Mind*.

Pietroski, P.M. 1994. Mental causation for dualists. *Mind and Language* 9:336-66.

Popper, K.R. 1953. Language and the body-mind problem: A restatement of interactionism. In *Proceedings of the 11th International Congress of Philosophy*. Reprinted in *Conjectures and the Growth of Scientific Knowledge*. Basic Books, 1962.

Popper, K.R. 1955. A note on the body-mind problem. *Analysis* 15:131-35.

Popper, K.R. 1977. Natural selection and the emergence of mind.

Scheffler, I. 1950. The new dualism: Psychological and physical terms.
Journal of Philosophy.

Sellars, W. 1954. A note on Popper's argument for dualism. Analysis 15:23-24.

Sussman, A. 1981. Reflection on the chances for a scientific dualism. Journal
of Philosophy 78:95-118.

Dualism is an empty hypothesis. Everything must be matter, though we may
have to expand the notion of matter.

Richardson, R.C. 1982. The 'scandal' of Cartesian dualism. Mind 91:20-37.

van Rooijen, K. 1987. Interactionism and evolution: A critique of Popper.
British Journal for the Philosophy of Science 38:87-92.

3.3e Psychophysical Relations, Misc

Campbell, K. 1983. Abstract particulars and the philosophy of mind.
Australasian Journal of Philosophy 61:129-41.

Caston, V. 1997. Epiphenomenalisms ancient and modern. Philosophical Review
106:309-363.

Hedman, C.G. 1970. On correlating brain states with psychological states.
Australasian Journal of Philosophy 48:247-51.

Heil, J. 1992. _The Nature of True Minds_. Cambridge University Press.

Honderich, T. 1981. Psychophysical law-like connections and their problems.
Inquiry 24:277-303.

Defending lawlike connections between physical states & conscious occurrents.
Contra anomalous monism and identity theory for occurrents. But occurrents
may not be causally efficacious. Comments by Wilson/Sprigge/Mackie/Stich.

McGinn, C. 1978. Mental states, natural kinds and psychophysical laws.
Proceedings of the Aristotelian Society 52:195-220. Reprinted in _The Problem
of Consciousness_ (Blackwell, 1991).

Argues that mental kinds are not natural kinds, and don't have real essences
but nominal essences. For this reason, there are no psychophysical laws.
With remarks on psychological laws, and the role of behavior.

Schechtman, M. 1997. The brain/body problem. Philosophical Psychology
10:149-64.

Scheerer, E. 1994. Psychoneural isomorphism: Historical background and current
relevance. Philosophical Psychology 7:183-210.

Skillen, A. 1984. Mind and matter: a problem which refuses dissolution. Mind
93:514-26.

Physical completeness, mental causation, non-reductionism are inconsistent.
Ryle and Putnam are closet dualists, and Davidson's an epiphenomenalist.

Steward, H. 1997. _The Ontology of Mind: Events, Processes, and States_.
Oxford University Press.

Tye, M. 1989. _The Metaphysics of Mind_. Cambridge University Press.

van Gelder, T. 1998. Monism, dualism, pluralism. Mind and Language 13:76-97.

3.4 Functionalism [see also 1.8, 4.6]

3.4a Causal Role Functionalism (Armstrong/Lewis)

Armstrong, D.M. 1968. A Materialist Theory of the Mind. Routledge and Kegan Paul.

Mental states should be analyzed as states that are apt to bring about certain kinds of behavior. Analysis of all kinds of mental states as such. With comments on dualism, behaviorism, identity theory, and consciousness.

Armstrong, D.M. 1970. The nature of mind. In (C. Borst, ed) The Mind/Brain Identity Theory. Macmillan. Reprinted in (N. Block, ed) Readings in the Philosophy of Psychology (MIT Press, 1980).

Mental states are internal states that are apt to cause certain behaviors. A synthesis between the "thesis" of idealism and the "antithesis" of behaviorism. With defense against objections from consciousness.

Clark, A. 1986. Psychofunctionalism and chauvinism. *Philosophy of Science* 53:535-59.

Psychofunctionalism can evade chauvinism by specifying different functional identifications within each species. Applying same mental terms to each is justified by theory similarity; but it still isn't analytic functionalism.

Goldstein, I. 1994. Identifying mental states: A celebrated hypothesis refuted. *Australasian Journal of Philosophy* 72:46-62.

Against functionalism: experiences have intrinsic introspectible acausal properties, such as duration, felt location, and unpleasantness. Both analytic and empirical functionalism fail.

Horgan, T. 1984. Functionalism and token physicalism. *Synthese* 59:321-38.

Formalizing versions of functionalism, and seeing which entail token physicalism and/or type physicalism. On the most plausible versions, we have token physicalism without type physicalism.

Hornsby, J. 1984. On functionalism, and on Jackson, Pargetter, and Prior on functionalism. *Philosophical Studies* 46:75-96.

Jackson, F., Pargetter, R. & Prior, E.W. 1982. Functionalism and type-type identity theories. *Philosophical Studies* 42:209-25.

Functionalism is compatible with type identity, as e.g. "pain" designates the state-type that fills the right functional role in an organism at a given time, i.e. a brain state. Contra Kripke, pain is not a rigid designator.

Kernohan, A. 1990. Lewis's functionalism and reductive materialism. *Philosophical Psychology* 3:235-46.

Argues that Lewis's functionalism founders on the specification of behavior. Described intentionally => non-materialist; physically => chauvinist.

Lewis, D. 1966. An argument for the identity theory. *Journal of Philosophy* 63:17-25. Reprinted in Philosophical Papers, Vol. 1 (Oxford University Press, 1980).

Causal roles are definitive of mental states. Since physical states fill these causal roles (by the explanatory adequacy of physics), mental states are physical states.

Lewis, D. 1972. Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50:249-58. Reprinted in (N. Block, ed) Readings in the Philosophy of Psychology (MIT Press, 1980).

Mental states can be defined, via a Ramsey-sentence analysis of the platitudes of folk psychology, as entities that fill causal roles specified by the analysis. These fillers turn out to be physical.

Lewis, D. 1978. Mad pain and martian pain. In (N. Block, ed) Readings in the Philosophy of Psychology, Vol. 1. MIT Press.

Accounting for both pains that don't play the usual causal role and for pains that are realized in different substances, by a mixed theory: pain is the physical state that typically occupies a certain causal role in a population.

McGinn, C. 1980. Functionalism and phenomenalism: A critical note. *Australasian Journal of Philosophy* 58:35-46. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Functionalism (reducing the mental to its effects on the physical) is no more plausible than phenomenalism (reducing the physical to its effects on the mental).

Owens, J. 1982. The failure of Lewis's functionalism. *Philosophical Quarterly* 36:159-73.

Lewis's original theory leads to Kripkean reference-fixing, so chauvinism. Token functionalism can't deal with paralytics. Species-relative functionalism fails as pain is intrinsic, not extrinsic.

Rogler, E. 2000. On David Lewis' philosophy of mind. *Protosociology* 14:285-311.

Sayward, C. 1995. Taking actions seriously. *Behavior and Philosophy* 23:51-60.

Shoemaker, S. 1981. Some varieties of functionalism. *Philosophical Topics* 12:93-119. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Fleshing out Ramsey-sentence functionalism; against Lewis's "mad pain" mixed theory; relating functionalism to the causal theory of properties. Empirical functionalism is chauvinistic so probably false. A terrific, in-depth paper.

Tye, M. 1983. Functionalism and type physicalism. *Philosophical Studies* 44:161-74.

Contra Lewis: Functionalism isn't compatible with type physicalism. There are intra-population difficulties with species-relative construals, and individual-relative construals can still have multiple fillers.

3.4b Machine Functionalism (Putnam) [see also 4.8]

Putnam, H. 1960. Minds and machines. In (S. Hook, ed) *Dimensions of Mind*. New York University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

The relationship between mental and physical states is just like that between logical and structural states of Turing Machines, so no great mystery. With comments on privacy and semantic analysis.

Putnam, H. 1967. The nature of mental states. In (Capitan & Merrill, eds) *Art, Mind, and Religion*. Pittsburgh University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Why mental states are more likely to be functional states (in probabilistic automata) than brain states or behavioral dispositions.

Putnam, H. 1967. The mental life of some machines. In (H. Castaneda, ed) *Intentionality, Minds and Perception*. Wayne State University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

On explaining behavior via TM states, e.g. explaining preference via utility functions. Logical behaviorism assumes rational preference functions. Functional organization is what matters, not physical make-up.

Putnam, H. 1975. Philosophy and our mental life. In *Mind, Language, and Reality*. Cambridge University Press.

Psychological states aren't TM states after all: we have lots of psych states at once; they depend on learning/memory; disjunctions of TM states are no

good. But functional organization rather than physics is still what counts.

Putnam, H. 1987. Representation and Reality. MIT Press.

Type functionalism isn't any better than type physicalism, as mental states can be multiply realized as functional states. With what in common?

Lycan, W.G. 1974. Mental states and Putnam's functionalist hypothesis.

Australasian Journal of Philosophy 52:48-62.

On abstract vs. physical TMs: Putnam should say that mental states are physical TM states. But then functionalism is compatible with physicalism. On the relation between Putnam's and Armstrong's functionalism.

Lycan, W.G. 1979. A New Lilliputian argument against machine functionalism.

Philosophical Studies 35:279-87.

If machine functionalism were true, a homunculus-head would have all the mental states of its homunculus (by the definition of "realization"), which is absurd.

Lycan, W.G. 1983. The moral of the New Lilliputian argument. Philosophical Studies 43:277-80.

Reply to Elugardo 1983: so how do you specify what count as inputs/outputs?

Elugardo, R. 1981. Machine functionalism and the New Lilliputian argument.

Pacific Philosophical Quarterly 62:256-61.

Criticism of Lycan 1979, and a re-making of the argument.

Elugardo, R. 1983. Machine realization and the New Lilliputian argument.

Philosophical Studies 43:267-75.

Lycan's New Lilliputian argument fails as inputs/outputs for the homunculus are not the same as inputs/outputs for the full system.

Kane, R.H. 1966. Turing machines and mental objects

Nelson, R. 1974. Mechanism, functionalism, and the identity theory. Journal of Philosophy 73:365-86.

Argues for mechanism rather than functionalism. Criticizes Putnam for hypostasizing mental states, which are disanalogous to mental states. Defending mechanism against Kalke's & Rorty's objections.

Rorty, R. 1972. Functionalism, machines and incorrigibility. Journal of Philosophy 69:203-20.

Logical states don't give us any understanding of mind over and above what the function/structure distinction gives us. In particular, it doesn't help with the understanding of privacy and incorrigibility.

Tomberlin, J. 1965. About the identity theory. Australasian Journal of Philosophy 43:295-99.

Contra Putnam: logical states are not physical states, and utterances about them are not about physical states.

Wagner, S.J. 1988. The liberal and the lycanthrope. Pacific Philosophical Quarterly 69:165-74.

Contra Lycan: machine functionalism can handle Bolivia and CRT cases by a causal/counterfactual account, and Lilliputian case by assigning mental states to minds, not bodies.

3.4c Functionalism, Miscellaneous

Adams, F. 1979. Properties, functionalism, and the identity theory. Eidos 1:153-79.

Bealer, G. 1978. An inconsistency in functionalism. Synthese.

A formal argument showing that functional definitions are equivalent to behavioral definitions.

Bealer, G. 1985. Mind and anti-mind: Why thinking has no functional definition. *Midwest Studies in Philosophy* 9:283-328.

Bechtel, W. 1984. Autonomous psychology: What it should and should not entail. *Philosophy of Science Association* 1984, 1:43-55.

The functional level is the appropriate level for psychology, but neurophysiological facts constrain this level and are thus relevant.

Ben-Yami, H. 1999. An argument against functionalism. *Australasian Journal of Philosophy* 77:320-324.

Biro, J.I. & Shahan, R.W. (eds) 1982. *Mind, Brain and Function*. Oklahoma University Press.

Ten papers on functionalism. Originally was *Philosophical Topics*, volume 12.

Block, N. 1980. Functionalism. In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol. 1. MIT Press.

Distinguishes varieties of functionalism, e.g. machine and Ramsey-sentence functionalism; and compares to behaviorism. With a historical overview, and arguments for why functionalism is incompatible with physicalism.

Block, N. 1978. Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9:261-325. Reprinted in *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Distinguishes analytic and empirical functionalism. Both have problems with absent qualia, and inputs/outputs. Analytic functionalism has problems with paralytics, etc; empirical functionalism has problems with Martians.

Block, N. & Fodor, J.A. 1972. What psychological states are not. *Philosophical Review* 81:159-81. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Mental states are not physical or behavioral states; could they be functional states? With various arguments against type identity, and against machine-table functionalism.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72:741-64. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

On the role of functional explanation versus other kinds of explanation. Functionalism applies an analytic, not subsumptive strategy.

David, M. 1997. Kim's functionalism. *Philosophical Perspectives* 11:133-48.

Fischer, J. 1985. Functionalism and propositions. *Philosophical Studies* 48:295-311.

Fodor, J.A. 1968. Materialism. In *Psychological Explanation*. Random House.

On mental state as inferred theoretical entities, individuated according to their function (cf. valve-lifters). Psychology and neuroscience will mutually constrain each other, giving a relation more complex than reduction.

Gendron, B. 1970. On the relation of neurological and psychological theories: A critique of the hardware thesis. *Boston Studies in the Philosophy of Science* 8:483-95.

Argues that functional explanation are reducible to structural explanations.

Hornsby, J. 1986. Physicalist thinking and conceptions of behaviour. In (P. Pettit & J. McDowell, eds) *Subject, Thought, and Context*. Oxford University Press.

Hoy, R.C. 1980. Dispositions, logical states, and mental occurrents. *Synthese* 44:207-40.n

Kalke, W. 1969. What's wrong with Fodor's and Putnam's functionalism. *Nous* 3:83-93.

There's no absolute functional/structural distinction, as it depends on how you choose boundaries and levels of abstraction.

Lycan, W.G. 1981. Form, function and feel. *Journal of Philosophy* 78:24-50.
Pursue a multi-leveled homuncular functionalism, with mental states characterized as states of teleologically identified subsystems. Even the identity theorist is a functionalist at a low level.

Malcolm, N. 1980. 'Functionalism' in philosophical psychology. *Proceedings of the Aristotelian Society* 80:211-30.

Pereboom, D. 1991. Why a scientific realist cannot be a functionalist. *Synthese* 88:341-58.

Scientific realism requires dispositions of kinds be explained by intrinsic properties. Neural/functional properties won't work, because of reductionism and circularity. Use intrinsic psychological properties instead.

Richardson, R.C. 1979. Functionalism and reductionism. *Philosophy of Science* 46:533-58.

Argues that functionalism is compatible with reductionism, by analogies. Genetics has multiple realization and multiple function; reduction doesn't require biconditionals. With remarks on the de facto autonomy of psychology.

Schiffer, S. 1986. Functionalism and belief. In (M. Brand & R. Harnish, eds) *The Representation of Knowledge and Belief*. University of Arizona Press.

Against functionalism for beliefs. Both common-sense functionalism and psychofunctionalism have problems with finding the right functional theory, distinguishing beliefs, perceptual input conditions, Twin Earth, etc.

Shope, R.K. 1973. Functional equivalence and the defense of materialism. *Philosophical Forum* 4:500-12.

Sober, E. 1990. Putting the function back into functionalism. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Need teleological functionalism, not Turing Machine functionalism.

Sober, E. 1985. Panglossian functionalism and the philosophy of mind. *Synthese* 64:165-93.

van Gulick, R. 1982. Functionalism as a theory of mind. *Philosophy Research Archives* 185-204.

The structure/function distinction is level-relative, so physiology might be relevant even under functionalism. Problems with automata, and with causal connections to nonintentionally characterized behavior.

van Gulick, R. 1980. Functionalism, information and content. *Nature and System* 2:139-62. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Ward, A. 1989. Philosophical functionalism. *Behaviorism* 17:155-8.

Weckert, J. 1990. Functionalism's impotence. *Philosophical Inquiry* 32-43.

Wilkes, K.V. 1981. Functionalism, psychology and the philosophy of mind. *Philosophical Topics* 12:147-67.

Functionalism may be appropriate for cognitive psychology but not for folk psychology, due to differing goals. Neuroscience will play an important role in developing functional theories.

Zangwill, N. 1992. Variable realization: not proven. *Philosophical Quarterly* 42:214-19.

Argues that the possibility of multiple realization has not been established, whether by arguments from imagination, concepts, or empirical facts.

3.5 Other Psychophysical Theories

3.5a Logical Behaviorism (Ryle, etc)

Ryle, G. 1949. *The Concept of Mind*. Hutchinson and Co.

The ancestor of most contemporary philosophy of mind. Among other things, argues that the "ghost in the machine" view of mind is a category mistake, and presents dispositional analyses of many mental concepts.

Bestor, T.W. 1979. Gilbert Ryle and the adverbial theory of mind. *Personalist* 60:233-42.

Campbell, C.A. 1953. Ryle on the intellect. *Philosophical Quarterly* 3:115-38.

Carnap, R. 1959. Psychology in physical language. In (Ayer, ed) *Logical Positivism*. Free Press.

Carrier, L. 1973. Professor Shaffer's refutation of behaviourism. *Mind* 80:249-52.

Chisholm, R. 1955. A note on Carnap's meaning analysis. *Philosophical Studies*.

Chisholm, R. 1952. Intentionality and the theory of signs. *Philosophical Studies*.

Chisholm, R. 1958. Sentences about believing. *Minnesota Studies in the Philosophy of Science* 2.

Dalrymple, H. 1977. Some logical muddles in behaviorism. *Southwestern Philosophical Studies* 2:64-72.

Ewing, A.C. 1953. Professor Ryle's attack on dualism. *Proceedings of the Aristotelian Society* 53:47-78.

Farrell, B. 1950. Experience. *Mind* 59:170-98.

Finn, D.R. 1971. Putnam and logical behaviourism. *Mind* 80:432-36.

Flanagan, O.J. & McCreadie-Albright, T. 1974. Malcolm and the fallacy of behaviorism. *Philosophical Studies* 26:425-30.

Geach, P. 1957. *Mental Acts*. Routledge and Kegan Paul.

Goudge, T.A. 1982. Ryle's last thoughts on thinking. *Dialogue* 21:125-32.

Graham, G. 19xx. Spartans and behaviorists. *Behaviorism* x:xx.

Defends behaviorism as a scientific hypothesis, so that conceivability arguments aren't relevant, and advocates "penetrability" behaviorism which can appeal to internal physical states.

Hamer, C. 1970. Why Ryle is not a behaviourist. *Philosophical Studies (Ireland)* 17:7-25.

Hamlyn, D.W. 1953. Behaviour. *Philosophy* 28:132-45.

Hanson, N.R. 1952. Professor Ryle's "mind". *Philosophical Quarterly* 2:246-48.

Heidelberger, H. 1966. On characterizing the psychological. *Philosophy and Phenomenological Research*.

Kitchener, R.F. 1977. Behavior and behaviorism. *Behaviorism* 5:11-68.

Jacquette, D. 1985. Logical behaviorism and the simulation of mental episodes. *Journal of Mind and Behavior* 6:325-332.

Mace, C.A. 1949. Some implications of analytical behaviourism. *Proceedings of the Aristotelian Society*.

Malcolm, N. 1954. Wittgenstein's *Philosophical Investigations*. *Philosophical Review* 43:530-9.

Mandelbaum, M. 1958. Professor Ryle and psychology. *Philosophical Review* 67:522-30.

McLaughlin, B. & O'Leary-Hawthorne, J. 1995. Dennett's logical behaviorism. *Philosophical Topics* 22:189-258.

Miller, D.S. 1911. Is consciousness "a type of behaviour"? *Journal of Philosophy* 8:322-27.

Miller, D.S. 1951. "Descartes myth" and "Professor Ryle's fallacy". *Journal of Philosophy*.

Nelson, R. 1969. Behaviorism is false. *Journal of Philosophy* 66:417-52.

Nelson, R. 1975. Behaviorism, finite automata, and stimulus-response theory. *Theory and Decision* 6:249-67.

Oosthuizen, D.C.S. 1970. Phenomenological psychology. *Mind* 79:487-501.

Park, S. 1994. Reinterpreting Ryle: A nonbehaviorist analysis. *Journal of the History of Philosophy* 32:265-90.

Place, U.T. 1993. A radical behaviorist methodology for the empirical investigation of private events. *Behavior and Philosophy* 20:25-35.

Price, H.H. 1960. Some objections to behaviorism. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

Putnam, H. 1963. Brains and behavior. In (R. Butler, ed) *Analytical Philosophy: Second Series*. Blackwell. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Quine, W.V. 1975. Mind and verbal dispositions. In (Guttenplan, ed) *Mind and Language*. Oxford University Press.

Quine, W.V. 1980. Sellars on behaviorism, language, and meaning. *Pacific Philosophical Quarterly* 61:26-30.

Robinson, H. 1982. Behaviorism and stimulus materialism. In *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Rowlands, M. 1991. A defense of behaviorism. *Behavior and Philosophy* 19:93-100.

Ryle, G. 1979. *On Thinking*. Blackwell.

Scriven, M. 1956. A study of radical behaviorism. *Minnesota Studies in the Philosophy of Science* 1:88-130.

Sellars, W. 1952. Mind, meaning, and behavior. Philosophical Studies.

Shuford, H. 1966. Logical behaviorism and intentionality. Theoria 32:246-51.

Skinner, B.F. 1945. The operational analysis of psychological terms. Psychological Review 52:270-78.

Smart, J.J.C. 1959. Ryle on mechanism and psychology. Philosophical Quarterly 9:349-55.

Stemmer, N. 1993. Behavioral materialism, the success of folk psychology, and the first-person case. Behavior and Philosophy 20:1-14.

Vendler, Z. 1981. Ryle's thoughts on thinking. Midwest Studies of Philosophy 6:335-43.

Weitz, M. 1951. Professor Ryle's "logical behaviourism". Journal of Philosophy 48:297-300.

Whitely, C.A. 1961. Behaviourism. Mind 70:164-74.

Wisdom, J. 1950. The concept of mind. Proceedings of the Aristotelian Society 50:189-204.

Wittgenstein, L. 1953. _Philosophical Investigations_.

Wright, J.N. 1959. Mind and the concept of mind. Aristotelian Society Supplement 33:1-22.

Ziff, P. 1958. About behaviourism. Analysis 18:132-6.

3.5b Identity Theory (Smart, etc) [see also 1.3c, 1.4g, 3.3b]

Feigl, H. 1958. The `mental' and the `physical'. Minnesota Studies in the Philosophy of Science 2:370-497. Reprinted as _The `Mental' and the `Physical'_. University of Minnesota Press, 1967.

Place, U.T. 1956. Is consciousness a brain process? British Journal of Psychology 47:44-50. Reprinted in (W. Lycan, ed) _Mind and Cognition_ (Blackwell, 1990).

The idea that consciousness is a brain process is logically coherent. It's a scientific hypothesis, not a necessary truth. On the "is" of composition vs the "is" of definition, and the fallacy of the internal phenomenal field.

Smart, J.J.C. 1959. Sensations and brain processes. Philosophical Review 68:141-56.

Defending the thesis that sensations are contingently identical to brain processes against various objections. Topic-neutral analysis of sensation reports. Materialism beats epiphenomenalism on grounds of simplicity.

Abelson, R. 1970. A refutation of mind-body identity. Philosophical Studies 18:85-90.

The number of possible mental states is infinite (think of any number), whereas there are only finitely many brain states, so they're not identical.

Armstrong, D.M. 1968. The headless woman and the defense of materialism. Analysis 29:48-49.

Likens the anti-materialist position to the "headless woman" fallacy: "I'm not aware the mental states are physical", so "I'm aware that mental states are non-physical".

Armstrong, D.M. 1973. Epistemological foundations for a materialist theory of

mind. *Philosophy of Science* 40:178-93.

A prima facie case for materialism based on grounds of rational consensus, arising especially from common-sense and scientific evidence. Mental states exist (common-sense) but should be analyzed causally (evidence from science).

Aune, B. 1966. Feigl on the mind-body problem. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Baier, K. 1962. Smart on sensations. *Australasian Journal of Philosophy* 40:57-68.

Mental states are necessarily private, and so cannot be physical states, which are public. We have epistemological authority about our mental states.

Beloff, J. 1965. The identity hypothesis: A critique. In (J.R. Smythies, ed) *Brain and Mind*. Routledge and Kegan Paul.

Blumenfeld, J-B. 1979. Phenomenal properties and the identity theory. *Australasian Journal of Philosophy* 63:485-93.

Argues that phenomenal properties aren't needed to identify sensations with brain-states, and nor are topic-neutral analyses.

Borst, C.V. (ed) 1970. *The Mind/Brain Identity Theory*. Macmillan.
An anthology of central articles on the identity theory.

Bradley, M.C. 1963. Sensations, brain-processes, and colours. *Australasian Journal of Philosophy* 41:385-93.

Brandt, R. 1960. Doubts about the identity theory. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

Brandt, R. & Kim, J. 1967. The logic of the identity theory. *Journal of Philosophy* 66:515-537.

Arguing for an event-identity construal of the identity theory. Comparing the identity theory to the weaker "principle of simultaneous isomorphism". The only reason to accept the identity theory is ontological simplicity.

Brodbeck, M. 1966. Mental and physical: Identity versus sameness. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Candlish, S. 1970. Mind, brain, and identity. *Mind* 79:502-18.

Carney, J. 1971. The compatibility of mind-body identity with dualism. *Mind*.
Argues that the identity theory is compatible with linguistic dualism, as the mental and the physical may differ in intensional properties only.

Clarke, J. 1971. Mental structure and the identity theory. *Mind* 80:521-30.

Coburn, R. 1963. Shaffer on the identity of mental states and brain processes. *Journal of Philosophy* 60:89-92.

Coder, D. 1973. The fundamental error of central-state materialism. *American Philosophical Quarterly* 10:289-98.

On problems with theories that leave the nature of mind open a priori: how can we even understand the possibilities?

Cornman, J. 1962. The identity of mind and body. *Journal of Philosophy* 59:486-92.

Coburn, R.C. 1963. Shaffer on the identity of mental states and brain processes. *Journal of Philosophy* 60:89.

Location of mental states by convention (Shaffer 1961) won't work, as it (a) makes mental states public, and (b) conflicts with connections to behavior.

Crittenden, C. 1971. Ontology and mind-body identity. *Philosophical Forum* 2:251-70.

de Boer, R. 1976. Cartesian categories in mind-body identity theories. *Philosophical Forum* 7:139-58.

Double, R. 1981. Central state materialism. *Philosophical Studies (Ireland)* 28:229-37.

Enc, B. 1983. In defense of the identity theory. *Journal of Philosophy* 80:279-98.

Feigl, H. 1971. Some crucial issues of mind-body monism. *Synthese*.

Garnett, A.C. 1965. Body and mind: the identity thesis. *Australasian Journal of Philosophy* 43:77-81.

Grunbaum, A. 1972. Abelson on Feigl's mind-body identity thesis. *Philosophical Studies* 23:119-21.

Gustafson, D.F. 1963. On the identity theory. *Analysis* 24:30-32.

Hanratty, G. 1972. The identity theory of Herbert Feigl. *Philosophical Studies* 20:113-23.

Harris, E.E. 1966. The neural identity thesis and the person. *International Philosophical Quarterly* 6:515-37.

Hedman, C.G. 1970. On correlating brain states with psychological states. *Australasian Journal of Philosophy* 48:247-51.

Heil, J. 1970. Sensations, experiences, and brain processes. *Philosophy* 45:221-6.

Hinton, J.M. 1967. Illusions and identity. *Analysis* 27:65-76.

Hockutt, M. 1967. In defense of materialism. *Philosophy and Phenomenological Research* 27:366-85.

Hoffman, R. 1967. Malcolm and Smart on brain-mind identity. *Philosophy* 42:128-36.

Joske, W. 1960. Sensations and brain processes: A reply to Professor Smart. *Australasian Journal of Philosophy* 38:157-60.

On topic-neutral reports, after-images, and after-radishes. Such a report requires epistemic access to physical resemblance, which we don't have.

Kim, J. 1966. On the psycho-physical identity theory. *American Philosophical Quarterly* 3:227-35.

There's no empirical support for identity, over and above that for correlation; and unity of science gives no reason to accept identity. The only reason might be that of ontological simplicity.

Kim, J. 1972. Phenomenal properties, psychophysical laws and the identity theory. *Monist* 56:178-92.

Deal with phenomenal properties by allowing only mental events, and eliminating mental objects. Identity theories needn't suppose psychophysical laws. With defense against multiple realizability arguments.

Kitcher, P.S. 1982. Two versions of the identity theory. *Erkenntnis* 17:213-28.

Recasting the identity theory and functionalism, using Kripkean theories of reference, so mental states can refer to physiological or psychological states that we don't yet understand; and qualia problems are handled better.

Lewis, D. 1965. An argument for the identity theory. *Journal of Philosophy* 63:17-25. Reprinted in *Philosophical Papers*, Vol. 1 (Oxford University Press, 1980).

Mental states are defined by their causal roles. So, by the completeness of physics, they must be physical states.

Locke, D. 1971. Must a materialist pretend he's anaesthetized? *Philosophical Quarterly* 49:217-31.

On how materialism, as opposed to a double aspect view, can handle mental features -- by moving them into the world via a realist theory of perception. Remarks on identification of states. After-images, etc, cause problems.

Lockwood, M. 1984. Einstein and the identity theory. *Analysis*.

Using the special theory of relativity to show that if mental events have a temporal location, then they must have a spatial location.

Lubow, N. 1978. Mind-body identity and irreducible properties. *Philosophy Research Archives* 4:1240.

Luce, D.R. 1966. Mind-body identity and psycho-physical correlation. *Philosophy of Science* 17:1-7.

Malcolm, N. 1964. Scientific materialism and the identity theory. *Dialogue* 3:115-25.

The identity theory is meaningless, if identity is analyzed as spatiotemporal coincidence, as thoughts don't have location. Thoughts also require context. Even if identity holds, explaining brain doesn't imply explaining mind.

Macdonald, C. 1989. *Mind-Body Identity Theories*. Routledge.

Malcolm, N. 1964. Scientific materialism and the identity theory. *Dialogue*.

Margolis, J. 1965. Brain processes and sensations. *Theoria* 31:133-38.

Meehl, P. 1966. The compleat autocerebroscopist: A thought-experiment on Professor Feigl's mind-body identity thesis. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Mucciolo, L. 1974. The identity theory and criteria for the mental. *Philosophy and Phenomenological Research* 35:167-80.

Munsat, S. 1969. Could sensations be processes? *Mind* 78:247-51.

Sensations and processes have different logical type, so it is a priori impossible that they should be identical.

Nagel, T. 1965. Physicalism. *Philosophical Review* 74:339-56, 1965.

Noren, S.J. 1970. Identity, materialism, and the problem of the danglers. *Metaphilosophy* 4:318-44.

Noren, S.J. 1970. Smart's materialism: The identity thesis and translation. *Australasian Journal of Philosophy* 48:54-66.

Norton, R. 1964. On the identity of identity theories. *Analysis* 25:14-16.

Pepper, S. 1975. A split in the identity theory. In (C. Cheng, ed) *Philosophical Aspects of the Mind-Body Problem*. Hawaii University Press.

Pitcher, G. 1960. Sensations and brain processes: A reply to Professor Smart.

Australasian Journal of Philosophy 38:150-7.

Identity requires explanation to be accepted, but Smart doesn't provide this. But one can deny identity without claiming dualism -- e.g. a "duck-rabbit" theory of mind/brain. With remarks on the completeness of descriptions.

Place, U.T. 1960. Materialism as a scientific hypothesis. Philosophical Review 69:101-4.

Contra Smart 1959: Materialism is a scientific hypothesis, if we accept certain logical criteria for what a sensation is; otherwise it's just false.

Place, U.T. 1972. Sensations and processes: A reply to Munsat. Mind.

Place, U.T. 1988. Thirty years on -- Is consciousness still a brain process?

Australasian Journal of Philosophy 66:208-19.

Comparing contemporary materialism to Pace's 1956 variety. With remarks on whether the thesis is empirical or a priori, and on deciding the issue between materialism and epiphenomenalism.

Place, U.T. 1989. Low claim assertions. In (J. Heil, ed) _Cause, Mind, and Reality: Essays Honoring C. B. Martin_. Kluwer.

Discusses a paper of Martin's and the genesis of the identity theory, with a focus on 'public' and 'private logic' and topic-neutral descriptions.

Presley, C.P. (ed) 1967. _The Identity Theory of Mind_. University of Queensland Press.

Puccetti, R. 1978. The refutation of materialism. Canadian Journal of Philosophy 8:157-62.

The identity theory must be false, as pain centers in vitro will not be pains. With a reply by G. Pearce and a rejoinder.

Ripley, C. 1969. The identity theory and scientific hypotheses. Dialogue 2:308-10.

Robinson, H. 1982. The disappearance theory. In _Matter and Sense: A Critique of Contemporary Materialism_. Cambridge University Press.

Rosenbaum, S. 1977. The property objection and the principles of identity. Philosophical Studies 32.

Routley, R. & MacCrae, V. 1966. On the identity of sensations and physiological occurrences. American Philosophical Quarterly 3.

Schlagel, R.H. 1977. The mind-body identity impasse. American Philosophical Quarterly 14:231-37.

Scriven, M. 1966. The limitations of the identity theory. In (P. Feyerabend & G. Maxwell, eds) _Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl_. University of Minnesota Press.

On the identity theory as a linguistic proposal, compatible with dualism; epiphenomenalism and parallelism must be false, leaving interactionism.

Sellars, W. 1965. The identity approach to the mind-body problem. Review of Metaphysics 18:430-51.

Shaffer, J. 1961. Could mental states be brain processes? Journal of Philosophy 58:813-22.

Mental states don't have a location, and brain processes do; but we could stipulate a location for mental states. With remarks on possible relations between mental and physical features, states, and concepts.

Shaffer, J. 1963. Mental events and the brain. Journal of Philosophy 60:160-6.

We identify mental events by noticing mental features that must be nonphysical, but still might be empirically reducible. Against topic-neutral definitions, and with response to Coburn 1963 on location.

Simon, M.A. 1970. Materialism, mental language, and the mind-body identity. *Philosophy and Phenomenological Research* 30:514-32.

Smart, J.J.C. 1960. Sensations and brain processes: A rejoinder to Dr. Pitcher and Mr. Joske. *Australasian Journal of Philosophy* 38:252-54.

Smart, J.J.C. 1961. Further remarks on sensations and brain processes. *Philosophical Review*.

Reply to Stevenson 1960: There are no irreducible mental properties; they reduce to physical properties via topic-neutral definitions.

Smart, J.J.C. 1962. Brain processes and incorrigibility. *Australasian Journal of Philosophy* 40:68-70.

Reply to Baier 1962: epistemological authority is compatible with materialism. Mental state reports are not completely incorrigible, though.

Smart, J.J.C. 1963. Materialism. *Journal of Philosophy* 60:651-62.

Defending topic-neutral analyses of mental reports, and arguing against Wittgensteinian behaviorism via brain-in-vat examples. With remarks on the appeal of materialism and on compatibility with ordinary language.

Smart, J.J.C. 1965. The identity thesis: A reply to Professor Garrett. *Australasian Journal of Philosophy* 43:82-3.

Smart, J.J.C. 1972. Further thoughts on the identity theory. *Monist* 56:177-92.

On some problems for the identity theory arising from the intensionality of mental states and from the appeal to properties, and on how to modify the translation form of the theory without embracing the disappearance version.

Smythies, J.R. 1994. Requiem for the identity theory. *Inquiry* 37:311-29.

Sosa, E. 1965. Professor Malcolm on "Scientific materialism and the identity theory". *Dialogue* 4:422-23.

Stevenson, J.T. 1960. 'Sensations and brain processes': A reply to J.J.C. Smart. *Philosophical Review* 69:505-10.

Identity theory implies nomological danglers, due to the irreducibility of defining mental properties.

Stoutland, F. 1971. Ontological simplicity and the identity hypothesis. *Philosophy and Phenomenological Research*.

The identity thesis isn't ontologically simpler than dualism: we still need a dualism of properties, and explanatory danglers. Not much turns on the issue, except in teleological explanation.

Sosa, E. 1965. Professor Malcolm on 'Scientific materialism and the identity theory'. *Dialogue* 3:422-23.

Contra Malcolm 1965: explaining brain will explain mind, if the explanation is conjoined with the identity statement. With rejoinder from Malcolm.

Swartz, N. 1974. Can the theory of contingent identity between sensation-states and brain-states be made empirical? *Canadian Journal of Philosophy* 3:405-17.

Swinburne, R. 1993. Are mental events identical with brain events? *American Philosophical Quarterly* 19:173-181.

Property identity theses fail due to meaning differences, and event identity these fail due to a lack of entailment relations. Rebuts objections from

weaker identity criteria and analogies with scientific identification.

Taylor, C. 1967. Mind-body identity, a side issue? *Philosophical Review* 76:201-13.

Teichmann, J. 1967. The contingent identity of minds and brains. *Mind* 76:404-15.

Thalberg, I. 1978. A novel approach to mind-brain identity. *Philosophy of Science* 3:255-72.

Suggests a theory in which neural states are components of, but not identical to, overall psychological states. This can accommodate raw feels if necessary as a further component, but is mostly materialistic.

Thomson, J.J. 1969. The identity theory. In (S. Morgenbesser, P. Suppes, & M. White, eds) *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*. St. Martin's Press.

Tomberlin, J.E. 1965. About the identity theory. *Australasian Journal of Philosophy* 53:295-9.

Watkins, J.W.N. 1978. A basic difficulty in the mind-brain identity hypothesis. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Weismann, D. 1965. A note on the identity thesis. *Mind* 74:571-77.

Whitely, C.H. 1970. The mind-brain identity hypothesis. *Philosophical Quarterly* 20:193-99.

Wolfe, J. & Nathan, G.J. 1968. The identity theory as a scientific hypothesis. *Dialogue* 7:469-72.

Ziedins, R. 1971. Identification of characteristics of mental events with characteristics of brain events. *American Philosophical Quarterly* 8:13-23.

3.5c Eliminative Materialism (Rorty, Feyerabend) [see also 1.4d, 1.7c, 2.1c]

Austin, J.W. 1975. Rorty's materialism. *Auslegung* 3:20-28.

Bernstein, R. 1968. The challenge of scientific materialism. *International Philosophical Quarterly* 8:252-75.

Bush, E. 1974. Rorty revisited. *Philosophical Studies* 25:33-42.

Cam, P. 1978. "Rorty revisited", or "Rorty revised". *Philosophical Studies* 33:377-86.

Carter, W.R. 1974. On incorrigibility and eliminative materialism. *Philosophical Studies* 28:113-21.

Cornman, J. 1968. On the elimination of 'sensations' and sensations. *Review of Metaphysics* 22:15-35.

Donovan, C. 1978. Eliminative materialism reconsidered. *Canadian Journal of Philosophy* 8.

Doppelt, G. 1977. Incorrigibility, the mental, and materialism. *Philosophy Research Archives*.

Everitt, N. 1981. A problem for the eliminative materialist. *Mind* 90:428-34.

Everitt, N. 1983. How not to solve a problem for the eliminative materialist. *Mind* 92:590-92.

Feyerabend, P. 1963. Mental events and the brain. *Journal of Philosophy* 40:295-6. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990). Identity theory implies dualism, though its acceptance of mental properties. Instead we should eliminate talk of mental processes altogether, or redefine them in physiological terms.

Feyerabend, P. 1963. Materialism and the mind-body problem. *Review of Metaphysics* 17:49-67.

Globus, G. 1989. The strict identity theory of Schlick, Russell, Maxwell, and Feigl. In (M. Maxwell & C. Savage, eds) *Science, Mind, and Psychology: Essays in Honor of Grover Maxwell*. University Press of America.

Godow, R. 1976. Eliminative materialism and denotation. *Philosophy and Phenomenological Research* 36.

Goodman, R.B. 1974. A note on eliminative materialism. *Journal of Critical Analysis* 5:80-83.

Hiley, D.R. 1978. Is eliminative materialism materialistic? *Philosophy and Phenomenological Research* 38:325-37.

Hiley, D.R. 1980. The disappearance theory and the denotation argument. *Philosophical Studies* 37:307-20.

Lycan, W.G. & Pappas, G. 1972. What is eliminative materialism? *Australasian Journal of Philosophy* 50:149-59.

Lycan, W.G. 1976. Quine's materialism. *Philosophia* 6:101-30.

Quine, W.V. 1966. On mental entities. In *The Ways of Paradox*. Random House.

Richardson, R.C. 1981. Disappearance and the identity theory. *Canadian Journal of Philosophy* 11:473-85.

Rorty, R. 1965. Mind-body identity, privacy, and categories. *Review of Metaphysics* 19:24-54.

Rorty, R. 1970. Incorrigibility as the mark of the mental. *Journal of Philosophy*.

Rorty, R. 1970. In defense of eliminative materialism. *Review of Metaphysics* 24:112-21.

Rosenthal, D.M. 1980. Keeping matter in mind. *Midwest Studies in Philosophy* 5:295-322.

Savitt, S. 1974. Rorty's disappearance theory. *Philosophical Studies* 28:433-36.

Shirley, E.S. 1974. Rorty's "disappearance" version of the identity theory. *Philosophical Studies* 25:73-75.

Sikora, R.I. 1974. Rorty's mark of the mental and his disappearance theory. *Canadian Journal of Philosophy* 4:191-93.

Sikora, R.I. 1975. Rorty's new mark of the mental. *Analysis* 35:192-94.

Steiling, K. 1976. The elimination of sensations and the loss of philosophy. *Auslegung* 3:20-28.

3.5d Anomalous Monism (Davidson)

Davidson, D. 1970. Mental events. In (L. Foster & J. Swanson, eds) Experience and Theory. Humanities Press. Reprinted in Essays on Action and Events (Oxford University Press, 1980).

Arguing for anomalous monism: no strict psychophysical laws, no strict psychological laws, and token identity without type identity. Mental events can still cause, via subsumption under physical laws.

Davidson, D. 1973. The material mind. In (P. Suppes, ed) Logic, Methodology and the Philosophy of Science. North-Holland. Reprinted in Essays on Action and Events (Oxford University Press, 1980).

The psychological supervenes on the physical but is not reducible to it, because of the holistic nature of intentional attribution. So building a perfect physical model may not explain psychology.

Davidson, D. 1974. Psychology as philosophy. In (S. Brown, ed) Philosophy of Psychology. Harper & Row. Reprinted in Essays on Action and Events (Oxford University Press, 1980).

On the differing constitutive standards of mental and physical concepts. Attribution of mental concepts is holistic, and presupposes a background of rationality, etc. With examples from decision theory.

Davidson, D. 1980. Essays on Actions and Events. Oxford University Press. A collection of papers on action, causation and the philosophy of psychology.

Davidson, D. 1987. Problems in the explanation of action. In (P. Pettit, R. Sylvan, & J. Norman, eds) Metaphysics and Morality. Blackwell.

Remarks on how mental properties can explain action without strict laws. The mental is a conceptual, not an ontological category, governed by normative standards, and not reducible to the non-normative.

Davidson, D. 1992. Thinking causes. In (J. Heil & A. Mele, eds) Mental Causation. Oxford University Press.

Davidson, D. 1995. Laws and cause. *Dialectica* 49:263-79.

Antony, L. 1989. Anomalous monism and the problem of explanatory force. *Philosophical Review* 98:153-87.

Criticism of Davidson's argument for rational causation. Reasons must cause in virtue of their rational properties. Token identities can't exist, due to normativity. Quinean psychology can't yield rational explanations.

Bickle, J. 1992. Mental anomaly and the new mind-brain reductionism. *Philosophy of Science* 59:217-30.

Campbell, N. 1997. The standard objection to anomalous monism. *Australasian Journal of Philosophy* 75:373-82.

Campbell, N. 1998. Anomalous monism and the charge of epiphenomenalism. *Dialectica* 52:23-39.

Cheng, K. 1997. Davidson's action theory and epiphenomenalism. *Journal of Philosophical Research* 22:81-95.

Child, W. 1993. Anomalism, uncodifiability, and psychophysical relations. *Philosophical Review*.

Anomalism is compatible with supervenience, if it is construed as denying psychophysical laws useful for explaining behavior. It is incompatible with token identity, though. With much on the uncodifiability of rationality.

Cooper, W.E. 1980. Materialism and madness. *Philosophical Papers* 9:36-40.

Daniel, S.G. 1999. Why even Kim-style psychophysical laws are impossible. *Pacific Philosophical Quarterly* 80:225-237.

Elgin, C. 1980. Indeterminacy, underdetermination and the anomalous monism. *Synthese* 45:233-55.

Garrett, B. 1999. Davidson on causal relevance. *Ratio* 12:14-33.

Goldberg, B. 1977. A problem with anomalous monism. *Philosophical Studies* 32:175-80.

Davidson's argument equivocates on the term "physical": the physical events that mental events cause might not be subsumed under laws.

Hess, P. 1981. Actions, reasons and Humean causes. *Analysis* 41:77-81.

Anomalous monism implies that mental properties don't cause anything.

Honderich, T. 1982. The argument for anomalous monism. *Analysis* 42:59-64.

If anomalous monism is true, mental events may cause, but their mental properties aren't causally relevant.

Johnston, M. 1985. Why having a mind matters. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

Anomalous monism loses out to Australian materialism. It can't be a priori, it leads to exhaustive monism, it doesn't support a new view of free action, and it implies the causal irrelevance of the mental.

Kalderon, M.E. 1987. Epiphenomenalism and content. *Philosophical Studies* 52:71-90.

Davidson's view leads to epiphenomenalism about content, as it can't support the appropriate counterfactuals. Strong supervenience might be a way out, but that is inconsistent with anomalism.

Kernohan, A. 1985. Psychology: Autonomous or anomalous? *Dialogue* 24:427-42.

Kim, J. 1985. Psychophysical laws. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

How there can be psychophysical generalizations but no laws -- they might lack modal force. On the relation between psychophysical anomalism and psychological anomalism. Casting Davidson as a Kantian dualist.

Kim, J. 1993. Can supervenience and "non-strict laws" save anomalous monism? In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Klagge, J.C. 1990. Davidson's troubles with supervenience. *Synthese* 85:339-52.

Anomalous supervenience is consistent, at the cost of anti-realism about the mental. Supervenience is a constraint on interpretation, but needn't support counterfactuals as different interpretation schemes are possible,

Klee, R. 1992. Anomalous monism, ceteris paribus, and psychological explanation. *British Journal for the Philosophy of Science* 43:389-403.

Problems with holism and ceteris paribus laws aren't unique to psychology. One finds the same thing in the physical sciences. So rationality plays no special role, and psychological laws are as reasonable as physical laws.

Kuczynski, J.M. 1998. A proof of the partial anomalousness of the mental. *Southern Journal Of Philosophy* 36:491-504.

Latham, N. 1999. Davidson and Kim on psychophysical laws. *Synthese* 118:121-44.

LePore, E. & Loewer, B. 1987. Mind matters. *Journal of Philosophy* 630-42.

Anomalous monism is not committed to epiphenomenalism, as even non-strict laws can ground counterfactuals and so support the causal relevance of mental properties.

Lycan, W.G. 1981. Psychological laws. *Philosophical Topics* 12:9-38.
A functionalist defense against anomalous monism. Psychofunctional laws and psychological laws, though not psychophysical laws, may exist. Rebutting arguments from rationality, indeterminism, intensionality, etc.

McDowell, J. 1985. Functionalism and anomalous monism. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.
Against Loar's functionalist reductionism: it doesn't begin to capture the normative role of rationality or the subjectivity of the mental.

McLaughlin, B.P. 1985. Anomalous monism and the irreducibility of the mental. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.
A very thorough summary of Davidson's views. Highly recommended.

McLaughlin, B.P. & LePore, E. (eds) 1985. *Actions and Events*. Blackwell.
30 essays on Davidson.

McLaughlin, B.P. 1992. On Davidson's response to the charge of epiphenomenalism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.
Comments on Davidson 1992. Davidson can respond to critics accepting causal relevance of mental properties and still denying strict laws. Davidson misconstrues his critics' positions on supervenience.

Melchert, N. 1986. What's wrong with anomalous monism. *Journal of Philosophy* 80:265-74.
Davidson is concerned with intentional, not phenomenal states; and his characterization of these is just as physical states under a certain description. So he avoids epiphenomenalism (contra e.g. Honderich 1982).

Miller, A. 1993. Some anomalies in Kim's account of Davidson. *Southern Journal of Philosophy* 31:335-44.
Kim's version of Davidson's argument against psychophysical laws cannot work. Elucidating the notion of a constitutive principle.

Noren, S.J. 1979. Anomalous monism, events, and 'the mental'. *Philosophy and Phenomenological Research* 40:64-74.

Patterson, S.A. 1996. The anomalism of psychology. *Proceedings of the Aristotelian Society* 96:37-52.

Preyer, G. 2000. Primary reasons: From radical interpretation to a pure anomalism of the mental. *Protosociology* 14:158-179.

Rosenberg, A. 1985. Davidson's unintended attack on psychology. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.
Anomalous monism implies that there aren't even heteronomic psychological generalizations, as variables can't be independently measured.

Rowlands, M. 1990. Anomalism, supervenience, and Davidson on content-individuation. *Philosophia* 295-310.
Supervenience is compatible with anomalism: biconditional laws are ruled out by the disjunctive base, and the wideness of mental states rules out one-way psychophysical laws, as there's no single property in the base.

Seager, W.E. 1981. The anomalousness of the mental. *Southern Journal of Philosophy* 19:389-401.
Elucidating Davidson's argument, focusing on the argument against strict psychophysical laws. Generalizations involve disjunctive kinds

and so are heteronomic and not law-like.

Seager, W.E. 1991. Disjunctive laws and supervenience. *Analysis* 51:93-98.
Argues contra Kim that supervenience is compatible with anomalous monism: the
the disjunctive generalizations aren't lawlike, as they aren't confirmed by
their instances.

Smart, J.J.C. 1985. Davidson's minimal materialism. In (B. Vermazen &
M. Hintikka, eds) *Essays on Davidson*. Oxford University Press.
Some comments on holism, indeterminacy, anomalism, and materialism.

Smith, P. 1982. Bad news for anomalous monism? *Analysis* 42:220-4.
Response to Honderich 1982: physical events are individuated as mental states
by virtue of their causal role, so the mental is causally relevant.

Sosa, E. 1993. Davidson's thinking causes. In (J. Heil & A. Mele, eds)
Mental Causation. Oxford University Press.

Stanton, W.L. 1983. Supervenience and psychophysical law in anomalous monism.
Pacific Philosophical Quarterly 64:72-9.
Supervenience entails psychophysical principles, but this is compatible with
anomalous monism. On what constitutes a strict psychophysical law.

Suppes, P. 1985. Davidson's views on psychology as a science. In (B. Vermazen
& M. Hintikka, eds) *Essays on Davidson*. Oxford University Press.
Various: physics is indeterministic and intensional, animals have beliefs,
psychology has derived laws, and decision-theory doesn't need speech.

Tiffany, E.C. 2001. The rational character of belief and the argument for
mental anomalism. *Philosophical Studies* 103:258-314.

van Gulick, R. 1980. Rationality and the anomalous nature of the mental.
Philosophy Research Archives 7:1404.
Rationality constraints don't introduce an irreducibly normative element into
intentional attributions. Rationality serves as a condition of adequacy for
psychophysical theories, but it doesn't rule them out.

Vermazen, B. & Hintikka, M. (eds) 1985. *Essays on Davidson*. Oxford
University Press.
12 essays on Davidson, with replies.

Walsh, D.M. 1998. Wide content individualism. *Mind* 107:625-652.

Welshon, R. 1999. Anomalous monism and epiphenomenalism. *Pacific
Philosophical Quarterly* 80:103-120.

Yalowitz, S. 1997. Rationality and the argument for anomalous monism.
Philosophical Studies 87:235-58.

Yalowitz, S. 1998. Causation in the argument for anomalous monism. *Canadian
Journal of Philosophy* 28:183-226.

Zangwill, N. 1993. Supervenience and anomalous monism: Blackburn on Davidson.
Philosophical Studies 71:59-79.

3.6 Mental Causation [see also 2.2c]

Antony, L. 1991. The causal relevance of the mental. *Mind and Language*
6:295-327.

Audi, R. 1993. Mental causation: Sustaining and dynamic. In (J. Heil &
A. Mele, eds) *Mental Causation*. Oxford University Press.

- Baker, L.R. 1993. Metaphysics and mental causation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.
Mental causation is incompatible with strong supervenience and causal closure of physics, as we can't distinguish high-level causes from non-causes. So reject the metaphysics and make explanation prior to causation.
- Barrett, J. 1994. Rationalizing explanation and causally relevant mental properties. *Philosophical Studies* 74:77-102.
- Blackburn, S. 1991. Losing your mind: Physics, identity, and folk burglar prevention. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.
Arguing for the causal efficacy and scientific respectability of higher-order states, such as functional-role states. To require appeal to particular physical states is to succumb to a "Tractarian" view of physical primacy.
- Block, N. 1989. Can the mind change the world? In (G. Boolos, ed) *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.
Rescuing content from epiphenomenalism via functional role argument; but then functional roles aren't really causally efficacious (cf. dormitive virtue), so epi all over again? Roles vs fillers, causation vs explanation.
- Block, N. 1995. Reply: Causation and two kinds of laws. In (C. Macdonald & G. Macdonald, eds) *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford University Press.
- Braun, D. 1995. Causally relevant properties. *Philosophical Perspectives* 9:447-75.
- Brewer, B. 1995. Compulsion by reason (Mental Causation II). *Aristotelian Society Supplement* 69:237-53.
- Burge, T. 1993. Mind-body causation and explanatory practice. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.
Mental causation is not a real worry, but the to-do shows that materialist metaphysics has shed little light on it. It needs to be understood at the mental level. With remarks on exclusion arguments and token identity.
- Crane, T. 1990. On an alleged analogy between numbers and propositions. *Analysis* 50:224-30.
How can a relation to a proposition (an abstract object) be causally efficacious? Analogy with numbers doesn't work: weight properties are only pseudo-relational, depending on units, but propositions are absolute.
- Crane, T. 1992. Mental causation and mental reality. *Proceedings of the Aristotelian Society* 66:185-202.
Argues that anomalism and causal closure don't pose problems for mental causation as they are false, and that functional properties can be efficacious. States with content may be efficacious, although content itself may not be.
- Crane, T. 1995. The mental causation debate (Mental causation I). *Aristotelian Society Supplement* 69:211-36.
Argues that mental causation is a deep problem for constitutive (but not identity) forms of physicalism. The only way out is to argue that it is a different variety of causation. But then what motivates physicalism?
- Dretske, F. 1993. Mental events as structuring causes of behavior. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.
Mental events are structuring causes of behavior; biological events are triggering causes, dependent on previous mental structuring. This allows extrinsic properties to play a causal role.

Ehring, D. 1996. Mental causation, determinables, and property instances. *Nous* 30:461-80.

Hardcastle, V.G. 1998. On the matter of minds and mental causation. *Philosophy and Phenomenological Research* 58:1-25.

Heil, J. 1992. Mentality and causality. *Topoi* 11:103-110.
On various problems with mental causation, and the relationship between psychology and philosophy.

Honderich, T. 1993. The union theory and anti-individualism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.
The identity theory and psychoneural correlation can't handle mental causation; only the union theory can. Anti-individualism causes problems, but should be rejected in any case.

Horgan, T. 1989. Mental causation. *Philosophical Perspectives* 3:47-74.
How mental events are causally relevant qua mental: via an account of "qua" causation in general, using counterfactuals on "pertinently similar worlds".

Horgan, T. 1997. Kim on mental causation and causal exclusion. *Philosophical Perspectives* 11:165-84.

Hornsby, J. 1993. Agency and causal explanation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Jackson, F. & Pettit, P. 1990. Causation and the philosophy of mind. *Philosophy and Phenomenological Research Supplement* 50:195-214.
A defense of functional role as a causally efficacious property of physical states. With application to connectionism & eliminativism.

Jackson, F. & Pettit, P. 1990. Program explanation: A general perspective. *Analysis* 50:107-17.

Jackson, F. 1995. Essentialism, mental properties, and causation. *Proceedings of the Aristotelian Society*.
How can content properties be causes, given that content is a matter of functional role and that functional properties are not causes? Defends a type-identity answer against various objections.

Jackson, F. 1996. Mental causation. *Mind* 105:377-413.
A "state of the art" review paper, concentrating on problems posed by autonomy, functionalism, and externalism, and advocating a sort of identity theory. With discussion of a "map-system" view vs. a language of thought.

Kazez, J.R. 1995. Can counterfactuals save mental causation? *Australasian Journal of Philosophy* 73:71-90.

Kim, J. 1984. Epiphenomenal and supervenient causation. *Midwest Studies in Philosophy* 9:257-70. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).
Psychological causation, like all macrocausation, is supervenient epiphenomenal causation.

Kim, J. 1992. The nonreductivist's trouble with mental causation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).
Argues that nonreductive materialism implies downward causation (as the mental has more causal powers than the physical alone), and that downward causation violates the causal closure of the physical.

Kim, J. 1992. "Downward causation" in emergentism and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or*

Reduction?: Prospects for Nonreductive Physicalism_. De Gruyter.

Argues that nonreductive materialism is just like 1930s emergentism, with the the mental contributing new causal powers, and so implies downward causation.

Kim, J. 1993. Mental causation in a physical world. In (E. Villanueva, ed) _Science and Knowledge_. Ridgeview.

Kim, J. 1994. `Second-order' properties and mental causation. Manuscript.

Kim, J. 1995. Mental causation: What? Me worry? In (E. Villanueva, ed) _Content_. Ridgeview.

Leiter, B. & Miller, A. 1994. Mind doesn't matter yet. Australasian Journal of Philosophy 72:220-28.

Argues that the arguments of Fodor and LePore & Loewer don't succeed in defeating the threat of epiphenomenalism.

Leiter, B. & Miller, A. 1998. Closet dualism and mental causation. Canadian Journal of Philosophy 28:161-181.

LePore, E. & Loewer, B. 1989. More on making mind matter. Philosophical Topics 17:175-91.

On the problems that irreducibility -- multiple realizability, normativity, and non-supervenience -- poses for mental causation. Criticizes Kim's supervenient causation and Fodor's causal powers, and looks to "quasation".

Macdonald, C. & Macdonald, G. 1986. Mental causes and explanation of action. Philosophical Quarterly 36:145-58.

Macdonald, C. & Macdonald, G. 1991. Mental causation and nonreductive monism. Analysis 51:23-32.

Macdonald, C. & Macdonald, G. 1995. How to be psychologically relevant. In (C. Macdonald & G. Macdonald, eds) _Philosophy of Psychology: Debates on Psychological Explanation_. Oxford University Press.

Macdonald, G. 1992. The nature of naturalism. Aristotelian Society Supplement 66:225-44.

Marras, A. 1994. Nonreductive materialism and mental causation. Canadian Journal of Philosophy 24:465-93.

Marras, A. 1997. The causal relevance of mental properties. Philosophia 25:389-400.

Marras, A. 1998. Kim's principle of explanatory exclusion. Australasian Journal of Philosophy 76:439-451.

McGrath, M. 1998. Proportionality and mental causation: A fit? Philosophical Perspectives 12:167-176.

McLaughlin, B.P. 1989. Type epiphenomenalism, type dualism, and the causal priority of the physical. Philosophical Perspectives 3:109-135.

Physical comprehensiveness and mental/physical non-reductionism don't imply mental inefficacy; nor does anomalous monism. Non-physical types can still can be causal, though they must be accompanied by physical causation.

Noordhof, P. 1997. Making the change: The functionalist's way. British Journal for the Philosophy of Science 48:233-??.

Noordhof, P. 1998. Do tropes resolve the problem of mental causation? Philosophical Quarterly 48:221-26.

Pettit, P. 1992. The nature of naturalism. Aristotelian Society Supplement

66:245-66.

On making sense of the causal efficacy of higher-level properties under naturalism. They're relevant at the program level, not quite in the way that basic properties are. With remarks on Macdonald's objections.

Robb, D. 1997. The properties of mental causation. *Philosophical Quarterly* 187:178-94.

Robb, D. 2001. Reply to Noordhof on mental causation. *Philosophical Quarterly* 51:90-94.

Robinson, W.S. 1979. Do pains make a difference to our behavior? *American Philosophical Quarterly* 16:327-34.

On Goldman's (1969) argument that dualism and causal closure are compatible with mental causation. Goldman establishes only hypothetical necessity, not causal necessity

Searle, J.R. 1984. Intentionality and its place in nature. *Synthese* 61:3-16. Intentionality is caused by the physical, and causes. More a 1P emphasis.

Sosa, E. 1984. Mind-body interaction and supervenient causation. *Midwest Studies in Philosophy* 9:271-81.

Interactionist dualism is out, supervenient causation is in. But there are problems with mental events' causal relevance qua mental, especially for anomalous monism. Cf: a loud shot causes death, but loudness isn't relevant.

Thomasson, A. 1998. A nonreductivist solution to mental causation. *Philosophical Studies* 89:181-95.

Tuomela, R. 1998. A defense of mental causation. *Philosophical Studies* 90:1-34.

van Gulick, R. 1993. Who's in charge here? And who's doing all the work? In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

On three arguments against mental causation, from strict laws, non-local supervenience, and especially exclusion. Mental properties are stable, recurring high-level patterns with their own causal relevance.

Worley, S. 1997. Determination and mental causation. *Erkenntnis* 46:281-304.

Yablo, S. 1992. Mental causation. *Philosophical Review* 101:245-280.

Argues that mental events/properties stand to physical events/properties as determinable to determinates, solving the exclusion problem. Some mental events are *better* candidates for the cause of action than physical events.

Zangwill, N. 1996. Good old supervenience: Mental causation on the cheap. *Synthese* 106:67-101.

Argues that anomalous monism is compatible with mental causation: supervenience is necessary and sufficient for causal efficacy.

3.7 Personal Identity

3.7a Personal Identity, General

Baillie, J. 1993. Recent work on personal identity. *Philosophical Books* 34:193-206.

Baillie, J. 1997. Personal identity and mental content. *Philosophical Psychology* 10:323-33.

Brennan, A. 1982. Personal identity and personal survival. *Analysis* 42:44-50.

- Brennan, A. 1984. Survival. Synthese 59:339-62.
- Brennan, A. 1987. Discontinuity and identity. Nous 21:241-60.
- Brennan, A. 1988. Conditions of Identity: A Study of Identity and Survival. Oxford University Press.
- Brooks, D.H.M. 1986. Group minds. Australasian Journal of Philosophy 64:456-70.
- Carter, W. 1999. Will I be a dead person? Philosophy and Phenomenological Research 59.
- Cartwright, H.M. 1987. Ruminations on an account of personal identity. In (J.J. Thomson, ed) On Being and Saying: Essays on Honor of Richard Cartwright. MIT Press.
- Cartwright, H.M. 1993. On two arguments for the indeterminacy of personal identity. Synthese 95:241-273.
- Cockburn, D. (ed) 1991. Human Beings. Cambridge University Press.
- Coleman, S. 2000. Thought experiments and personal identity. Philosophical Studies 98:51-66.
- Cowley, F. 1971. The identity of a person and his body. Journal of Philosophy 68:678-683.
- Dainton, B. 1996. Survival and experience. Proceedings of the Aristotelian Society 96:17-36.
- Davis, L.H. 1998. Functionalism and personal identity. Philosophy and Phenomenological Research 58:781-804.
- Davis, L.H. 2001. Functionalism, the brain, and personal identity. Philosophical Studies 102:259-79.
- Dennett, D.C. 1978. Where am I? In Brainstorms. MIT Press.
- Elliot, R. 1991. Personal identity and the causal continuity requirement. Philosophical Quarterly 41:55-75.
- Ganeri, J. 2000. Cross-modality and the self. Philosophy and Phenomenological Research 61:639-658.
- Garrett B. 1990. Personal identity and extrinsicness. Philosophical Studies 59:177-194.
- Garrett, B. 1991. Personal identity and reductionism. Philosophy and Phenomenological Research 51:361-373.
- Garrett, B. 1992. Persons and values. Philosophical Quarterly 42:337-44.
- Glover, J. 1988. I: The Philosophy and Psychology of Personal Identity. Penguin.
- Hamilton, A. 1995. A new look at personal identity. Philosophical Quarterly 45:332-349.
- Harris, H. (ed) 1995. Identity. Oxford University Press.
- Harris, H. 1995. An experimentalist looks at identity. In (H. Harris, ed) Identity. Oxford University Press.

- Hasker, W. 1999. *The Emergent Self*. Cornell University Press.
- Hope, T. 1994. Personal Identity and Psychiatric Illness. *Philosophy* 37:131-143.
- Johnston, M. 1992. Reasons and reductionism. *Philosophical Review* 3:589-618.
- Kolak, D. & Martin, R. 1987. Personal identity and causality: Becoming unglued. *American Philosophical Quarterly*.
- Kolak, D. 1993. The metaphysics and metapsychology of personal identity: Why thought experiments matter in deciding who we are. *American Philosophical Quarterly* 30:39-50.
- Kolak, D. & Martin, R. (eds) 1991. *Self and Identity: Contemporary Philosophical Issues*. Macmillan.
- Madell, G. 1981. *The Identity of the Self*. Edinburgh University Press.
- Madell, G. 1991. Personal identity and the idea of a human being. *Philosophy* 29:127-142.
- Martin, R. 1992. Self-interest and survival. *American Philosophical Quarterly* 29:319-30.
- Matthews, S. 2000. Survival and separation. *Philosophical Studies* 98:279-303.
- McCall, C. 1990. *Concepts of Person: An Analysis of Concepts of Person, Self, and Human Being*. Avebury.
- Merricks, T. 2000. Perdurantism and psychological continuity. *Philosophy and Phenomenological Research* 61:195-199.
- Miri, M. 1973. Memory and personal identity. *Mind* 82:1-21.
- Nerlich, G.C. 1958. Sameness, difference, and continuity. *Analysis*.
- Noonan, H. 1989. *Personal Identity*. Routledge.
- Noonan, H. 1993. Chisholm, persons, and identity. *Philosophical Studies* 69:35-58.
- Nozick, R. 1981. The identity of the self. In *Philosophical Explanations*. Harvard University Press.
- Olson E. 1994. Is Psychology relevant to personal identity? *Australasian Journal of Philosophy* 72:173-186.
- Olson, E.T. 1997. *The Human Animal: Personal Identity without Psychology*. Oxford University Press.
- Olson, E. 2001. Personal identity and the radiation argument. *Analysis* 61:38-44.
- Peacocke, A. & Gillett, G. (eds) 1987. *Persons and Personality: A Contemporary Inquiry*. Blackwell.
- Penelhum, T. 1959. Personal identity, memory, and survival. *Journal of Philosophy*.
- Penelhum, T. 1971. The importance of self-identity. *Journal of Philosophy* 68:667-78.
- Perry, J. 1972. Can the self divide? *Journal of Philosophy* 69:463-88.

- Perry, J. (ed) 1975. *Personal Identity*. University of California Press.
- Perry, J. 1975. Personal identity, memory, and the problem of circularity. In (J. Perry, ed) *Personal Identity*. University of California Press.
- Perry, J. 1976. The importance of being identical. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Perry, J. 1978. *A Dialogue on Personal Identity and Immortality*. Hackett.
- Pogue, J.E. 1993. Identity, survival, and the reasonableness of replication. *Southern Journal of Philosophy* 31:45-70.
- Rea, M. & Silver, D. 2000. Personal identity and psychological continuity. *Philosophy and Phenomenological Research* 61:185-194.
- Rey, G. 1976. Survival. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Rieber, S. 1998. The concept of personal identity. *Philosophy and Phenomenological Research* 58:581-594.
- Robert, M. 1983. Lewis's theory of personal identity. *Australasian Journal of Philosophy* 61:58-67.
- Rorty, A. (ed) 1976. *The Identities of Persons*. University of California Press.
- Shalom, A. 1985. *The Body-Mind Conceptual Framework and the Problem of Personal Identity*. Humanities Press.
- Schechtman, M. 1990. Personhood and personal identity. *Journal of Philosophy* 87:71-92.
- Shoemaker, S. 1959. Personal identity and memory. *Journal of Philosophy* 56:868-902.
- Shoemaker, S. 1970. Persons and their pasts. *American Philosophical Quarterly* 7:269-85.
- Shoemaker, S. & Swinburne, S. 1984. *Personal Identity: Great Debates in Philosophy*. Blackwell.
- Shorter, J.M. 1962. More about bodily continuity and personal identity. *Analysis* 22:79-85.
- Unger, P. 1990. *Identity, Consciousness, and Value*. Oxford University Press.
- Vesey, P. 1974. *Personal Identity: A Philosophical Analysis*. Cornell University Press.
- White, S. 1989. Metapsychological relativism and the self. *Journal of Philosophy* 86:298-323.
- Whiting, J. 1986. Friends and future selves. *Philosophical Review* 95:547-80.
- Wilkes, K.V. 1988. *Real People: Personal Identity Without Thought Experiments*. Oxford University Press.
- Williams, B. 1957. Personal identity and individuation. *Proceedings of the Aristotelian Society* 67:229-52.

Williams, B. 1973. *_Problems of the Self_*. Cambridge University Press.

Zemach, E. 1987. Looking out for number one. *Philosophy and Phenomenological Research*.

Zuboff, A. 1978. Moment universals and personal identity. *Proceedings of the Aristotelian Society* 52:141-55.

Zuboff, A. 1990. One self: The logic of experience. *Inquiry* 33:39-68.

3.7b Parfit on Personal Identity

Baillie, J. 1993. What matters in survival. *Southern Journal of Philosophy* 31:255-61.

Baillie, J. 1996. Identity, relation R, and what matters: A challenge to Derek Parfit. *Pacific Philosophical Quarterly* 77:263-267.

Beck, S. 1989. Parfit and the Russians (personal identity and moral concepts). *Analysis* 49:205-209.

Bodansky, E. 1987. Parfit on selves and their interests. *Analysis* 47:47-50.

Brennan, A.A. 1987. Survival and importance. *Analysis* 47:225-30.

Brueckner, A. 1993. Parfit on what matters in survival. *Philosophical Studies* 70:1-22.

Bushnell, D.E. 1993. Identity, psychological continuity, and rationality. *Journal of Philosophical Research* 18:15-24.

Campbell, S. 2000. Strawson, Parfit and impersonality. *Canadian Journal of Philosophy* 30:207-225.

Cassam, Q. 1993. Parfit on persons. *Proceedings of the Aristotelian Society* 93:17-37.

Chappell, T. 1995. Personal identity, R-relatedness, and the empty question argument. *Philosophical Quarterly* 45:88-92.

Chappell, T. 1998. Reductionism about persons; and what matters. *Proceedings of the Aristotelian Society* 98:41-58.

Collins, A.W. 1997. Personal identity and the coherence of q-memory. *Philosophical Quarterly* 47:73-80.

Curzer, H. 1991. An ambiguity in Parfit's theory of personal identity. *Ratio* 4:16-24.

Dancy, J. (ed). 1997. *_Reading Parfit_*. Blackwell.

Doepke, F. 1990. The practical importance of personal identity. *Logos* 83-91.

Ehring, D. 1987. Survival and trivial facts. *Analysis* 47:50-54.

Ehring, D. 1995. Personal identity and the R-relation: Reconciliation through cohabitation. *Australasian Journal of Philosophy* 73:337-346.

Fields, L. 1987. Parfit on personal identity and desert. *Philosophical Quarterly* 37:432-41.

Gillett, G. 1987. Reasoning about persons. In (A. Peacocke & G. Gillett, eds) *_Persons and Personality: A Contemporary Inquiry_*. Blackwell.

- Goodenough, J.M. 1996. Parfit and the Sorites paradox. *Philosophical Studies* 2:113-20.
- Haugen, D. 1995. Personal identity and concern for the future. *Philosophia* 24:481-492.
- Hirsch, E. 1991. Divided minds. *Philosophical Review* 1:3-30.
- Johnston, M. 1989. Fission and the facts. *Philosophical Perspectives* 3:369-97.
- Korsgaard, C. 1989. Personal identity and the unity of agency: A Kantian response to Parfit. *Philosophy and Public Affairs* 18:103-31.
- Lee, W. 1990. Personal identity, the temporality of agency, and moral responsibility. *Auslegung* 16:17-29.
- Lewis, D. 1976. Survival and identity. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Madell, G. 1985. Derek Parfit and Greta Garbo. *Analysis* 45:105-9.
- Maddy, P. 1979. Is the importance of identity derivative? *Philosophical Studies* 35:151-70.
- Matthews, G.B. 1977. Surviving as. *Analysis* 37:53-58.
- Martin, R. 1987. Memory, connecting, and what matters in survival. *Australasian Journal of Philosophy* 65:82-97.
- Measor, N. 1980. On what matters in survival. *Mind* 89:406-11.
- Merricks, T. 1997. Fission and personal identity over time. *Philosophical Studies* 88:163-186.
- Northoff, G. 2000. Are "q-memories" empirically realistic?: A neurophilosophical approach. *Philosophical Psychology* 13:191-211.
- Oaklander, L.N. 1987. Parfit, circularity, and the unity of consciousness. *Mind* 96:525-29.
- Parfit, D. 1971. Personal identity. *Philosophical Review* 80:3-27.
- Parfit, D. 1971. On the importance of self-identity. *Journal of Philosophy* 68:683-90.
- Parfit, D. 1973. Later selves and moral principles. In (A. Montefiore, ed) *Philosophy and Personal Relations*. Routledge and Kegan Paul.
- Parfit, D. 1976. Lewis, Perry, and what matters. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Parfit, D. 1982. Personal identity and rationality. *Synthese* 53.
- Parfit, D. 1984. *Reasons and Persons*. Oxford University Press.
- Parfit, D. 1995. The unimportance of identity. In (H. Harris, ed) *Identity*. Oxford University Press.
- Robinson, J. 1988. Personal identity and survival. *Journal of Philosophy* 85:319-28.
- Rovane, C. 1990. Branching self-consciousness. *Philosophical Review* 99:355-95.

- Siderits, M. 1988. Ehring on Parfit's relation R. *Analysis* 48:29-32.
- Slors, M. 2001. Personal identity, memory, and circularity: An alternative for q-memory. *Journal of Philosophy* 98:186-214.
- Sprigge, T.L.S. 1988. Personal and impersonal identity. *Mind* 97:29-49.
- Storl, H. 1992. The problematic nature of parfitian persons. *Personalist Forum* 8:123-31.
- Stone, J. 1988. Parfit and the Buddha: Why there are no people. *Philosophy and Phenomenological Research* 48:519-32.
- Wolf, S. 1986. Self-interest and interest in selves. *Ethics* 96:704-20.

3.7c Persons

- Aune, B. 1994. Speaking of selves. *Philosophical Quarterly* 44:279-93.
- Barresi, J. 1999. On becoming a person. *Philosophical Psychology* 12:79-98.
- Bertocci, P.A. 1978. The essence of a person. *Monist* 61:28-41.
- Biro, J.I. 1981. Persons as corporate entities and corporations as persons. *Nature and System* 3:173-80.
- Chisholm, R.M. 1976. *Person and Object: A Metaphysical Study*. Open Court.
- Dennett, D.C. 1976. Conditions of personhood. Lewis, D. 1976. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Dennett, D.C. 1989. The origins of selves. *Cogito* 3:163-73.
- Heinimaa, M. 2000. Ambiguities in the psychiatric use of the concepts of the person: An analysis. *Philosophy, Psychiatry, and Psychology* 7:125-136.
- Lowe, E.J. 1991. Real selves: Persons as a substantial kind. *Philosophy* 29:87-107.
- Johnston, M. 1987. Human beings. *Journal of Philosophy* 84:59-83.
- Margolis, J. 1988. Minds, selves, and persons. *Topoi* 7:31-45.
- McInerney, P.K. 1998. Persons and psychological systems. *American Philosophical Quarterly* 35:179-193.
- McInerney, P.K. 2000. Conceptions of persons and persons through time. *American Philosophical Quarterly* 37:121-134.
- Oderberg, D. 1989. Johnston on human beings. *Journal of Philosophy* 86:137-41.
- Olson, E. 1998. Human atoms. *Australasian Journal of Philosophy* 76:396-406.
- Peterson, J. 1985. Persons and the problem of interaction. *Modern Schoolman* 62:131-38.
- Rorty, A.O. 1976. A literary postscript: Characters, persons, selves, individuals. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Smart, B. 1976. Synchronous and diachronous selves. *Canadian Journal of Philosophy* 6:13-33.

Strawson, P. 1958. Persons. Minnesota Studies in the Philosophy of Science 2:330-53.

Unger, P. 1979. I do not exist. In (G. Macdonald, ed) *Perception and Identity*. Cornell University Press.

Unger, P. 1979. Why there are no people. Midwest Studies in Philosophy 4:177-222.

Vincent, A. 1989. Can groups be persons? Review of Metaphysics 42:687-715.

Wiggins, D. 1987. The person as object of science, as subject of experience, and as locus of value. In (A. Peacocke & G. Gillett, eds) *Persons and Personality*. Blackwell.

3.7d Split Brains [see also 6.1e]

Baillie, J. 1991. Split brains and single minds. Journal of Philosophical Research 16:11-18.

Davis, L. 1997. Cerebral hemispheres. Philosophical Studies 87:207-22.

Gill, J.H. 1980. Of split brains and tacit knowing. International Philosophical Quarterly 20:49-58.

Gillett, G. 1986. Brain bisection and personal identity. Mind 95:224-9.

Greenwood, J.D. 1993. Split brains and singular personhood. Southern Journal of Philosophy 31:285-306.

Marks, C. 1980. *Commissurotomy, Consciousness, and Unity of Mind*. MIT Press.

Martin, R. 1995. Fission rejuvenation. Philosophical Studies 80:17-40.

Merricks, T. 1997. Fission and personal identity over time. Philosophical Studies 88:163-186.

Mills E. 1993. Dividing without reducing: Bodily fission and personal identity. Mind 102:37-51.

Moor, J. 1982. Split brains and atomic persons. Philosophy of Science 49:91-106.

Nagel, T. 1971. Brain bisection and the unity of consciousness. Synthese 22:396-413. Reprinted in *Mortal Questions* (Cambridge University Press, 1979).

Parfit, D. 1987. Divided minds and the nature of persons. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Puccetti, R. 1973. Brain bisection and personal identity. British Journal for the Philosophy of Science 24:339-55.

Puccetti, R. 1973. Multiple identity. Personalist 54:203-13.

Puccetti, R. 1975. The mute self: A reaction to DeWitt's alternative account of the split-brain data. British Journal for the Philosophy of Science 27:65-73.

Puccetti, R. 1981. The case for mental duality: Evidence from split-brain data and other considerations. Behavioral and Brain Sciences 4:93-123.

Puccetti, R. 1989. Two brains, two minds. *British Journal for the Philosophy of Science* 40:137-44.

Puccetti, R. 1993. Mind with a double brain. *British Journal for the Philosophy of Science* 44:675-92.

Puccetti, R. 1993. Dennett on the split-brain. *Psychology* 4(52).

Robinson, D.N. 1976. What sort of persons are hemispheres? Another look at "split-brain" man. *British Journal for the Philosophy of Science* 27:73-8.

Shaffer, J. 1977. Personal identity: The implications of brain bisection and brain transplants. *Journal of Medicine and Philosophy* 2:147-61.

Sperry, R.W. 1984. Consciousness, personal identity and the divided brain. *Neuropsychologia* 22:611-73.

3.7e Multiple Personality

Apter, A. 1991. The problem of who: Multiple personality, personal identity, and the double brain. *Philosophical Psychology* 4:219-48.

Benner, D.G., Evans, C.S. 1984. Unity and multiplicity in hypnosis, commissurotomy, and multiple personality disorder. *Journal of Mind and Behavior* 5:423-431.

Boden, M.A. 1994. Multiple personality and computational models. *Philosophy* 37:103-114.

Braude, S.E. 1991. *_First-person Plural: Multiple Personality and the Philosophy of Mind_*. Routledge.

Braude, S.E. 1996. Multiple personality disorder and moral responsibility. *Philosophy, Psychiatry, and Psychology* 3:37-54.

Clark, S.R.L. 1991. How many selves make me? *Philosophy* 29:213-33.

Clark, S.R.L. 1996. Minds, memes, and multiples. *Philosophy, Psychiatry, and Psychology* 3:21-28.

Flanagan, O. 1994. Multiple identity, character transformation, and self-reclamation. In (G. Graham & G. Stephens, eds) *_Philosophical Psychopathology_*. MIT Press.

Gillett, G. 1986. Multiple personality and the concept of a person. *New Ideas in Psychology* 4:173-84.

Gillett, G. 1997. A discursive account of multiple personality disorder. *Philosophy, Psychiatry, and Psychology* 4:213-22.

Hacking, I. 1991. Two souls in one body. *Critical Inquiry* 17:838-67.

Hacking, I. 1995. *_Rewriting the Soul: Multiple Personality and the Sciences of Memory_*. Princeton University Press.

Humphrey, N. & Dennett, D.C. 1989. Speaking for ourselves. *Raritan* 9:68-98.

Kolak, D. 1993. Finding our selves: Identification, identity, and multiple personality. *Philosophical Psychology* 6:363-86.

Lizza, J.P. 1993. Multiple personality and personal identity revisited. *British Journal for the Philosophy of Science* 44:263-274.

Matthews, S. 1998. Personal identity, multiple personality disorder, and moral personhood. *Philosophical Psychology* 11:67-88.

Radden, J. 1996. *_Divided Minds and Successive Selves: Ethical Issues in Disorders of Identity and Personality_*. MIT Press.

Wilkes, K.V. 1981. Multiple personalty and personal identity. *British Journal for the Philosophy of Science* 32:331-48.

Wilkes, K.V. 1991. How many selves make me? *Philosophy* 66:235-43.

3.8 Free Will

Albritton, R. 1985. Freedom of the will and freedom of action. *Proceedings and Addresses of the American Philosophical Association* 59:239-51.

Anscombe, G.E.M. 1976. 'Soft' determinism. In (G. Ryle, ed) *_Contemporary Aspects of Philosophy_*. Oriel Press.

Audi, R. 1974. Moral responsibility, freedom, and compulsion. *American Philosophical Quarterly* 11:1-14.

Ayer, A.J. 1980. Free will and rationality. In (Z. van Straaten, ed) *_Philosophical Subjects_*. Oxford University Press.

Ayers, M. 1968. *_The Refutation of Determinism_*. Methuen.

Berofsky, B. (ed) 1966. *_Free Will and Determinism_*. Harper and Row.

Berofsky, B. 1971. *_Determinism_*. Princeton University Press.

Berofsky, B. 1987. *_Freedom from Necessity: The Metaphysical Basis of Responsibility_*. Routledge.

Bishop, J. 1993. Compatibilism and the free will defense. *Australasian Journal of Philosophy* 71:104-20.

Blumenfeld, D. 1971. The principle of alternate possibilities. *Journal of Philosophy* 67:339-44.

Blumenfeld, D. 1988. Freedom and mind control. *American Philosophical Quarterly* 25:215-27.

Campbell, C.A. 1951. Is "free will" a pseudoproblem? *Mind* 60:441-65.

Churchland, P.S. 1981. Is determinism self-refuting? *Mind* 90:99-101.

Clarke, R. 1992. Free will and the conditions of moral responsibility. *Philosophical Studies* 66:53-72.

Clarke, R. 1993. Toward a credible agent-causal account of free will. *Nous* 27:191-203.

Clarke, R. 2000. Modest libertarianism. *Philosophical Perspectives* 14:21-46.

Crisp, T. & Warfield, T. 2000. The irrelevance of indeterministic counterexamples to principle beta. *Philosophy & Phenomenological Research* 61:173-185.

Dennett, D.C. 1984. *_Elbow Room: The Varieties of Free Will Worth Wanting_*. MIT Press.

Double, R. 1989. Puppeteers, hypnotists, and neurosurgeons. *Philosophical*

Studies 56:163-73.

Double, R. 1991. *The Non-Reality of Free Will*. Oxford University Press.

Double, R. 1991. Determinism and the experience of freedom. *Pacific Philosophical Quarterly* 72:1-8.

Double, R. 1992. How rational must free will be? *Metaphilosophy* 23:268-78.

Double, R. 1994. How to frame the free will problem. *Philosophical Studies* 75:149-72.

Double, R. 1996. *Metaphilosophy and Free Will*. Oxford University Press.

Duggan, T. & Gert, B. 1979. Free will as the ability to will. *Nous* 13:197-217.

Dworkin, G. (ed) 1970. *Determinism, Free Will, and Moral Responsibility*. Prentice-Hall.

Eccles, J. 1976. Brain and free will. In (G. Globus, ed) *Consciousness and the Brain*. Plenum Press.

Fischer, J.M. 1982. Responsibility and control. *Journal of Philosophy* 79:24-40.

Fischer, J.M. & Ravizza, M. 1992. When the will is free. *Philosophical Perspectives* 6:423-51.

Fischer, J.M. 1994. *The Metaphysics of Free Will*. Blackwell.

Fischer, J.M. & Ravizza, M. 1996. Free will and the modal principle. *Philosophical Studies* 3:213-30.

Fowler, C. 1996. A pragmatic defense of free will. *Journal of Value Inquiry* 30:247-60.

Frankfurt, H. 1969. Alternate possibilities and moral responsibility. *Journal of Philosophy* 65:829-39.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68:5-20.

Furlong, F.W. 1981. Determinism and free will: Review of the literature. *American Journal of Psychiatry* 138:435-39.

Garson, J.W. 1995. Chaos and free will. *Philosophical Psychology* 8:365-74.

Ginet, C. 1980. The conditional analysis of freedom. In (P. van Inwagen, ed) *Time and Cause: Essays Presented to Richard Taylor*. Reidel.

Ginet, C. 1983. In defense of incompatibilism. *Philosophical Studies* 44:391-400.

Ginet, C. 1989. Reasons explanation of action: An incompatibilist account. *Philosophical Perspectives* 3.

Ginet, C. 1990. *On Action*. Cambridge University Press.

Goldman, A. 1968. Actions, predictions, and books of life. *American Philosophical Quarterly*.

Goldman, A. 1969. The compatibility of mechanism and purpose. *Philosophical Review* 78:468-82.

- Griffiths, A.P. 1989. Is free will incompatible with something or other? *Philosophy* 24:101-19.
- Greenspan, P.S. 1978. Behavior control and freedom of action. *Philosophical Review* 87:225-40.
- Greenspan, P.S. 1993. Free will and the genome project. *Philosophy and Public Affairs* 22:31-43.
- Heller, M. 1996. The mad scientist meets the robot cats: Compatibilism, kinds, and counterexamples. *Philosophy and Phenomenological Research* 56:333-37.
- Honderich, T. (ed) 1973. *Essays on Freedom of Action*. Routledge and Kegan Paul.
- Honderich, T. 1988. *A Theory of Determinism*. Oxford University Press.
- Honderich, T. 1993. *How Free Are You?* Oxford University Press.
- Horgan, T. 1985. Compatibilism and the consequence argument. *Philosophical Studies* 47:339-56.
- Hospers, J. 1950. Meaning and free will. *Philosophy and Phenomenological Research* 10:307-30.
- Howard, G.S. 1993. Steps toward a science of free will. *Counseling and Values* 37:116-28.
- Kane, R. 1985. *Free Will and Values*. SUNY Press.
- Kane, R. 1989. Two kinds of incompatibilism. *Philosophy and Phenomenological Research* 69:219-54.
- Kane, R. 1994. Free will: The illusive ideal. *Philosophical Studies* 75:25-60.
- Kane, R. 1996. *The Significance of Free Will*. Oxford University Press.
- Kane, R. 2000. The dual regress of free will and the role of alternative possibilities. *Philosophical Perspectives* 14:57-80.
- Kapitan, T. 1986. Deliberation and the presumption of open alternatives. *Southern Journal of Philosophy* 40:230-51.
- Kapitan, T. 1991. Ability and cognition: A defense of compatibilism. *Philosophical Studies* 63:231-43.
- Kapitan, T. 1996. Modal principles in the metaphysics of free will. *Philosophical Perspectives* 10:419-45.
- Kapitan, T. 2000. Autonomy and manipulated freedom. *Philosophical Perspectives* 14:81-104.
- Kenny, A. 1976. *Will, Freedom, and Power*. Blackwell.
- Kenny, A. 1978. *Free Will and Responsibility*. Routledge.
- Klein, M. 1990. Determinism, blameworthiness, and deprivation. Oxford University Press.
- Ladd, J. 1952. Free will and voluntary action. *Philosophy and Phenomenological Research* 12:392-405.
- Lahav, R. 1991. Between pre-determinism and arbitrariness: A Bergsonian approach to free will. *Southern Journal of Philosophy* 29:487-99.

- Lamb, J.W. 1977. On a proof of incompatibilism. *Philosophical Review* 86:20-35.
- Lamb, J.W. 1993. Evaluative compatibilism and the principle of alternate possibilities. *Journal of Philosophy* 90:517-27.
- Lehrer, K. (ed) 1966. *Freedom and Determinism*. Random House.
- Lehrer, K. 1966. An empirical disproof of determinism. In (K. Lehrer, ed) *Freedom and Determinism*. Random House.
- Lehrer, K. 1976. 'Can' in theory and practice: A possible worlds analysis. In (M. Brand & D. Walton, eds) *Action Theory*. Reidel.
- Locke, D. 1975. Three concepts of free action. *Aristotelian Society Supplement* 75:95-112.
- Lucas, J.R. 1970. *The Freedom of the Will*. Oxford University Press.
- Machina, K. 1994. Challenges for compatibilism. *American Philosophical Quarterly* 31:213-22.
- Margenau, H. 1931. The uncertainty principle and free will. *Science*.
- McCall, S. 1984. Freedom defined as the power to decide. *American Philosophical Quarterly* 21:329-38.
- Morden, M. 1990. Free will, self-causation, and strange loops. *Australasian Journal of Philosophy* 68:59-73.
- Morgenbesser, S. & Walsh, J.J. (eds) 1962. *Freedom and Responsibility*. Prentice-Hall.
- Narveson, J. 1977. Compatibilism defended. *Philosophical Studies* 32:83-7.
- Neely, W. 1974. Freedom and desire. *Philosophical Review* 83:32-54.
- O'Connor, D.J. 1971. *Free Will*. Anchor Books.
- O'Connor, T. 1993. Indeterminism and free agency: Three recent views. *Philosophy and Phenomenological Research* 53:499-26.
- O'Connor, T. (ed) 1995. *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. Oxford University Press.
- O'Leary-Hawthorne, J. & Pettit, P. 1996. Strategies for free will compatibilists. *Analysis* 56:191-201.
- O'Shaughnessy, B. 1980. *The Will: A Dual Aspect Theory*. Cambridge University Press.
- Pereboom, D. 2000. Alternative possibilities and causal histories. *Philosophical Perspectives* 14:119-138.
- Perszyk, K.J. 1999. Compatibilism and the free will defence: A reply to Bishop. *Australasian Journal of Philosophy* 77:92-105.
- Popper, K. 1983. Is determinism self-refuting? *Mind* 92:103-4.
- Ravizza, M. 1994. Semi-compatibilism and the transfer of non-responsibility. *Philosophical Studies* 75:61-93.
- Rowe, W. 1987. Two concepts of freedom. *Proceedings and Addresses of the*

American Philosophical Association 61:43-64.

Rychlak, J.F. 1976. Can psychology be objective about free will? *Philosophical Psychologist* 10:2-9. Revised version in *New Ideas in Psychology* 1:213-29, 1983.

Rychlak, J.F. 1994. Four kinds of determinism and "free will": A response to Viney and Crosby. *New Ideas in Psychology* 12:143-46.

Rychlak, J.F. 1994. Is free will a process or a content: Both? neither? Are we free to take a position on this question? *Journal of Theoretical and Philosophical Psychology* 14:62-72.

Sappington, A.A. 1990. Recent psychological approaches to the free will versus determinism controversy. *Psychological Bulletin* 108:19-29.

Searle, J. 2000. Consciousness, free action and the brain. *Journal of Consciousness Studies* 7:3-22.

Settle, T. 1993. How determinism refutes compatibilism. *Religious Studies* 29:353-62.

Slife, B.D. 1994. Free will and time: That "stuck" feeling. *Journal of Theoretical and Philosophical Psychology* 14:1-12.

Slote, M.A. 1969. Free will, determinism, and the theory of important criteria. *Inquiry* 12:317-38.

Slote, M. 1980. Understanding free will. *Journal of Philosophy* 77:136-51.

Slote, M. 1982. Selective necessity and the free will problem. *Journal of Philosophy* 74:5-24.

Spence, S.A. 1996. Free will in the light of neuropsychiatry. *Philosophy, Psychiatry, and Psychology* 3:75-90.

Stampe, D.W. & Gibson, M.I. 1992. Of one's own free will. *Philosophy and Phenomenological Research* 52:529-56.

Strawson, G. 1986. *Freedom and Belief*. Oxford University Press.

Stump, E. & Fischer, J. 2000. Transfer principles and moral responsibility. *Philosophical Perspectives* 14:47-56.

Thorp, J. 1980. *Free Will: A Defense Against Neurophysiological Determinism*. Routledge.

van Inwagen, P. 1975. The incompatibility of free will and determinism. *Philosophical Studies* 27:185-99.

van Inwagen, P. 1978. Ability and responsibility. *Philosophical Review* 87:201-24.

van Inwagen, P. 1983. *An Essay on Free Will*. Oxford University Press.

van Inwagen, P. 1989. When is the will free? *Philosophical Perspectives* 3.

van Inwagen, P. 1994. When the will is not free. *Philosophical Studies* 75:95-113.

Van Inwagen, P. 2000. Free will remains a mystery. *Philosophical Perspectives* 14:1-20.

- Vesey, G. 1989. Responsibility and free will. *Philosophy* 24:85-100.
- Vihvelin, K. 2000. Freedom, foreknowledge, and the principle of alternate possibilities. *Canadian Journal of Philosophy* 30:1-23.
- Vihvelin, K. 2000. Libertarian compatibilism. *Philosophical Perspectives* 14:139-166.
- Viney, D.W. & Crosby, D.A. 1994. Free will in process perspective. *New Ideas in Psychology* 12:129-41.
- Waller, B.N. 1989. Uneven starts and just deserts (fatalism and free will). *Analysis* 49:209-13.
- Waller, B. 1990. *Freedom without Responsibility*. Temple University Press.
- Warfield, T. 2000. Causal determinism and human freedom are incompatible: A new argument for incompatibilism. *Philosophical Perspectives* 14:167-180.
- Watson, G. (ed) 1982. *Free Will*. Oxford University Press.
- Westcott, M.R. 1977. Free will: An exercise in metaphysical truth or psychological consequences. *Canadian Psychological Review* 18:249-63.
- Widerker, D. 2000. Frankfurt's attack on the principle of alternative possibilities: A further look. *Philosophical Perspectives* 14:181-202.
- Williams, C. 1980. *Free Will and Determinism: A Dialogue*. Hackett.
- Wilton, R. 2000. *Consciousness, Free Will, and the Explanation of Human Behavior*. E. Mellen Press.
- Wolf, S. 1980. Asymmetrical freedom. *Journal of Philosophy* 77:151-66.
- Wolf, S. 1981. The importance of free will. *Mind* 90:366-78.
- Wolf, S. 1990. *Freedom within Reason*. Oxford University Press.
- Yaffe, G. 2000. Free will and agency at its best. *Philosophical Perspectives* 14:203-230.
- Zagzebski, L. 2000. Does libertarian freedom require alternate possibilities? *Philosophical Perspectives* 14:231-248.

Zimmerman, D. 1994. Acts, omissions, and semi-compatibilism. *Philosophical Studies* 73:209-23.

3.9 The Problem of Other Minds

- Alexander, P. 1959. Other people's experiences. *Proceedings of the Aristotelian Society*.
- Aune, B. 1961. The problem of other minds. *Philosophical Review*.
- Austin, J. 1946. Other minds. *Aristotelian Society Supplement* 20:148-87.
- Ayer, A.J. 1953. One's knowledge of other minds. *Theoria*.
- Ayer, A.J. 1956. *The Problem of Knowledge*. Harmondsworth.
- Baron-Cohen, S., Tager-Flusberg, H., Cohen, D.J. 1994. *Understanding Other Minds: Perspectives from Autism*. Oxford University Press.
- Buck, R. 1962. Non-other minds. In (R. Butler, ed) *Analytic Philosophy*.

Barnes and Noble.

Buford, T.O. 1970. *Essays on Other Minds*. University of Illinois Press.

Castaneda, H. 1962. Criteria, analogy, and knowledge of other minds. *Journal of Philosophy*.

Duhrssen, A. 1963. Philosophic alienation and the problem of other minds. *Philosophical Review*.

Everett, T. 2000. Other voices, other minds. *Australasian Journal of Philosophy* 78:213-222.

Feigl, H. 1959. Other minds and the egocentric predicament. *Journal of Philosophy* 56:980-87.

Gallagher, K. 1964. Intersubjective knowledge. In (Sheed & Ward, eds) *The Philosophy of Knowledge*.

Glennan, S.S. 1995. Computationalism and the problem of other minds. *Philosophical Psychology* 8:375-88.

Hampshire, S. 1952. The analogy of feeling. *Mind* 61:1-12.

Harnad, S. 1991. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* 1:43-54.
On the Total Turing Test (full behavioral equivalence) as a test for mind.

Hyslop, A. 1976. Other minds as theoretical entities. *Australasian Journal of Philosophy* 54:158-61.

Hyslop, A. 1995. *Other Minds*. Kluwer.

Jones, J.R. 1950. Our knowledge of other persons. *Philosophy* 25.

Jorgensen, J. 1949. Remarks concerning the concept of mind and the problem of other people's minds. *Theoria*.

Kurthen, M. Moskopp, D., Linke, D.B. & Reuter, B.M. 1991. The locked-in syndrome and the behaviorist epistemology of other minds. *Theoretical Medicine* 12:69-79.

Lenman, J. 1994. Beliefs about other minds: A pragmatic justification. *American Philosophical Quarterly* 31:223-34.

Locke, D. 1968. *Myself and Others: A Study in our Knowledge of Minds*. Oxford University Press.

Malcolm, N. 1958. Knowledge of other minds. *Journal of Philosophy*.

Melnyk, A. 1994. Inference to the best explanation and other minds. *Australasian Journal of Philosophy* 4:482-91.

Mellor, W.W. 1956. Three problems about other minds. *Mind* 65:200-217.

Morick, H. (ed) 1967. *Wittgenstein and the Problem of Other Minds*. Humanities Press.

Narveson, A.H. 1966. Evidential necessity and other minds. *Mind* 75.

Pap, A. 1951. Other minds and the principle of verifiability. *Revue Internationale de Philosophie* 5:280-306.

Peacocke, C. 1984. Consciousness and other minds. *Aristotelian Society*

Supplement 58:97-117.

Plantinga, A. 1966. Induction and other minds. Review of Metaphysics 19:441-61.

Plantinga, A. 1967. God and Other Minds. Cornell University Press.

Plantinga, A. 1968. Induction and other minds II. Review of Metaphysics 12:524-33.

Price, H.H. 1938. Our evidence for the existence of other minds. Philosophy 13:425-56.

Sagal, P. & Borg, G. 1993. The range principle and the problem of other minds. British Journal for the Philosophy of Science 44:477-91.

Slote, M. 1966. Induction and other minds. Review of Metaphysics 20:341-60.

Sober, E. 2000. Evolution and the problem of other minds. Journal of Philosophy 97:365-387.

Spencer, W. 1930. Our Knowledge of Other Minds. Yale University Press.

Sprigge, T.L.S. 1992. Ayer on other minds. In (L. Hahn, ed) The Philosophy of A.J. Ayer. Open Court.

Thalberg, I. 1969. Other times, other places, other minds. Philosophical Studies 20.

Thomson, J.F. 1951. The argument from analogy and the problem of other minds. Mind 60:336-50.

Weinberg, J. 1946. Our knowledge of other minds. Philosophical Review 60.

Wisdom, J. 1946. Other minds. Aristotelian Society Supplement 20:122-47.

Wisdom, J. 1968. Other Minds. University of California Press.

Zemach, E. 1966. Sensations, raw feels, and other minds. Review of Metaphysics 20:317-40.

Ziff, P. 1965. The simplicity of other minds. Journal of Philosophy 42:575-84.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 4: Philosophy of Artificial Intelligence [565]

=====

- 4.1 Can Machines Think? [224]
 - 4.1a The Turing Test [51]
 - 4.1b Godelian Arguments (Lucas, Penrose) [59]
 - 4.1c The Chinese Room (Searle) [58]
 - 4.1d Machine Consciousness [33]
 - 4.1e Machine Thought, Misc [23]
- 4.2 Computation and Representation [44]
 - 4.2a Symbols and Symbol Systems [9]
 - 4.2b Computational Semantics [16]
 - 4.2c Implicit/Explicit Representation [8]
 - 4.2d AI without Representation? [6]
 - 4.2e Miscellaneous [6]
- 4.3 Philosophy of Connectionism [150]
 - 4.3a Connectionism and Compositionality (Fodor/Pylyshyn) [43]
 - 4.3b Representation in Connectionism [20]
 - 4.3c Connectionism and Eliminativism [16]
 - 4.3d The Connectionist/Classical Debate [20]
 - 4.3e Subsymbolic Computation (Smolensky) [8]
 - 4.3f Philosophy of Connectionism, Misc [36]
 - 4.3g Foundational Empirical Issues [12]
- 4.4 Dynamical Systems [22]
- 4.5 Foundational Questions in AI [51]
 - 4.5a The Nature of AI [7]
 - 4.5b Levels of Analysis (Marr, etc) [12]
 - 4.5c The Frame Problem [14]
 - 4.5d Methodology [19]
- 4.6 Computationalism in Cognitive Science [33]
- 4.7 Computation and Physical Systems [18]
- 4.8 Philosophy of AI, Misc [17]

4.1 Can Machines Think?

4.1a The Turing Test

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59:433-60.

Proposes the Imitation game (Turing test) as a test for intelligence: If a machine can't be told apart from a human in a conversation over a teletype, then that's good enough. With responses to various objections.

Alper, G. 1990. A psychoanalyst takes the Turing test. *Psychoanalytic Review* 77:59-68.

Barresi, J. 1987. Prospects for the Cyberiad: Certain limits on human self-knowledge in the cybernetic age. *Journal for the Theory of Social Behavior* 17:19-46.

Block, N. 1981. Psychologism and behaviorism. *Philosophical Review* 90:5-43.
A look-up table could pass the Turing test, and surely isn't intelligent.
The TT errs in testing behavior and not mechanisms. A nice, thorough paper.

Bringsjord, S. 2001. Creativity, the Turing test, and the (better) Lovelace test. *Minds & Machines* 11:3-27.

Clark, T. 1992. The Turing test as a novel form of hermeneutics. *International Studies in Philosophy* 24:17-31.

Copeland, B.J. 2000. The Turing test. *Minds and Machines* 10:519-539.

Crawford, C. 1994. Notes on the Turing test. Communications of the Association for Computing Machinery 37:13-15.

Crockett, L. 1994. _The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence_. Ablex.

Davidson, D. 1990. Turing's test. In (K. Said, ed) _Modelling the Mind_. Oxford University Press.

Dennett, D.C. 1984. Can machines think? In (M. Shafto, ed) _How We Know_. Harper & Row.

Defending the Turing test as a good test for intelligence.

Erion, G.J. 2001. The Cartesian test for automatism. Minds and Machines 1:29-39.

French, R.M. 1990. Subcognition and the limits of the Turing test. Mind 99:53-66.

The Turing Test is too hard, as it requires not intelligence but human intelligence. Any machine could be unmasked through careful questioning, but this wouldn't mean that the machine was unintelligent.

French, R.M. 1995. Refocusing the debate on the Turing Test: A response. Behavior and Philosophy 23:59-60.

Response to Jacquette 1993.

Gunderson, K. 1964. The imitation game. Mind 73:234-45.

The Turing test is not broad enough: there's much more to thought than the ability to play the imitation game.

Harnad, S. 1991. Other bodies, other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1:43-54.

On the Total Turing Test (full behavioral equivalence) as a test for mind.

Harnad, S. 1994. Levels of functional equivalence in reverse bioengineering: The Darwinian Turing test for artificial life. Artificial Life 1(3).

Harnad, S. 1999. Turing on reverse-engineering the mind. Journal of Logic, Language, and Information.

Hauser, L. 1993. Reaping the whirlwind: Reply to Harnad's "Other bodies, other minds". Minds and Machines 3:219-37.

Hauser, L. 2001. Look who's moving the goal posts now. Minds and Machines 11:41-51.

Hayes, P. & Ford, K. 1995. Turing test considered harmful. _Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence_ 1:972-77.

Hofstadter, D.R. 1981. A coffee-house conversation on the Turing test. Scientific American.

A dialogue on the Turing test.

Jacquette, D. 1993. Who's afraid of the Turing test? Behavior and Philosophy 20:63-74.

Defending the Turing test against French 1990. Turing did not intend the test to provide a *necessary* condition for intelligence.

Jacquette. D. 1993. A Turing test conversation. Philosophy 68:231-33.

Karelis, C. 1986. Reflections on the Turing test. Journal for the Theory of Social Behavior 16:161-72.

Lee, E.T. 1996. On the Turing test for artificial intelligence. *Kybernetes* 25:61.

Leiber, J. 1989. Shanon on the Turing test. *Journal of Social Behavior*.

Leiber, J. 1995. On Turing's Turing Test and why the matter matters. *Synthese* 104:59-69.

Turing's test is neutral about the structure of the machine that passes it, but it must be practical and reliable (thus excluding Searle's and Block's counterexamples).

Mays, W. 1952. Can machines think? *Philosophy* 27:148-62.

Michie, D. 1993. Turing's test and conscious thought. *Artificial Intelligence* 60:1-22. Reprinted in (P. Millican & A. Clark, eds) *_Machines and Thought_*. Oxford University Press.

Millar, P. 1973. On the point of the Imitation Game. *Mind* 82:595-97.

Moor, J.H. 1976. An analysis of Turing's test. *Philosophical Studies* 30:249-257.

The basis of the Turing test is not an operational definition of thinking, but rather an inference to the best explanation.

Moor, J.H. 1978. Explaining computer behavior. *Philosophical Studies* 34:325-7.

Reply to Stalker 1978: Mechanistic and mentalistic explanations are no more incompatible than program-based and physical explanations.

Moor, J.H. 2001. The status and future of the Turing test. *Minds and Machines* 11:77-93.

Piccinini, G. 2000. Turing's rules for the imitation game. *Minds and Machines* 10:573-582.

Purthill, R. 1971. Beating the imitation game. *Mind* 80:290-94.

Rankin, T.L. 1987. The Turing paradigm: A critical assessment. *Dialogue* 29:50-55.

Some obscure remarks on lying, imitation, and the Turing test.

Richardson, R.C. 1982. Turing tests for intelligence: Ned Block's defense of psychologism. *Philosophical Studies* 41:421-6.

A weak argument against Block: input/output function doesn't guarantee a capacity to respond sensibly.

Rosenberg, J. 1982. Conversation and intelligence. In (B. de Gelder, ed) *_Knowledge and Representation_*. Routledge & Kegan Paul.

Sampson, G. 1973. In defence of Turing. *Mind* 82:592-94.

Saygin, A.P., Cicekli, I. & Akman V. 2000. Turing test: 50 years later. *Minds and Machines* 10:463-518.

Schweizer, P. 1998. The truly total Turing Test. *Minds and Machines* 8:263-272.

Shanon, B. 1989. A simple comment regarding the Turing test. *Journal for the Theory of Social Behavior* 19:249-56.

The Turing test presupposes a representational/computational framework for cognition. Not all phenomena can be captured in teletype communication.

Shieber, S.M. 1994. Lessons from a restricted Turing test. *Communications of*

the Association for Computing Machinery 37:70-82.

Stalker, D.F. 1978. Why machines can't think: A reply to James Moor. *Philosophical Studies* 34:317-20.

Contra Moor 1976: The best explanation of computer behavior is mechanistic, not mentalistic.

Sterrett, S.G. 2000. Turing's two tests for intelligence. *Minds and Machines* 10:541-559.

Stevenson, J.G. 1976. On the imitation game. *Philosophia* 6:131-33.

Traiger, S. 2000. Making the right identification in the Turing test. *Minds and Machines* 10:561-572.

Watt, S. 1996. Naive psychology and the inverted Turing test. *Psychology* 7(14).

Whitby, B. 1996. The Turing test: AI's biggest blind alley? In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Zdenek, S. 2001. Passing Loebner's Turing test: A case of conflicting discourse functions. *Minds & Machines* 11:53-76.

4.1b Godelian arguments (Lucas, Penrose)

Benacerraf, P. 1967. God, the Devil, and Godel. *Monist* 51:9-32.

Discusses and sharpens Lucas's arguments. Argues that the real consequence is that if we are Turing machines, we can't know which.

Bowie, G. 1982. Lucas' number is finally up. *Journal of Philosophy Logic*, 11:279-85.

Lucas's very Godelization procedure makes him inconsistent, unless he has an independent way to see if any TM is consistent, which he doesn't.

Boyer, D. 1983. J.R. Lucas, Kurt Godel, and Fred Astaire. *Philosophical Quarterly* 33:147-59.

Remarks on the various ways in which Lucas and a machine might be said to "prove" anything, and the ways in which a machine might simulate Lucas. The argument has all sorts of level confusions, and a bit of circularity.

Chari, C. 1963. Further comments on minds, machines and Godel. *Philosophy* 38:175-8.

Can't reduce the lawless creative process to computation.

Chalmers, D.J. 1996. Minds, machines, and mathematics. *Psyche* 2:11-20.

Chihara, C. 1972. On alleged refutations of mechanism using Godel's incompleteness results. *Journal of Philosophy* 64:507-26.

An analysis of the Lucas/Benacerraf argument. On various senses in which a machine might come to know its own program.

Coder, D. 1969. Godel's theorem and mechanism. *Philosophy* 44:234-7.

Only mathematicians understand Godel, so Lucas's argument isn't general; and Turing machines can go wrong. Weak.

Dennett, D.C. 1978. The abilities of men and machines. In *Brainstorms*. MIT Press.

There is no unique TM which we are -- there could be many.

Edis, T. 1998. How Godel's theorem supports the possibility of machine intelligence. *Minds and Machines* 8:251-262.

Feferman, S. 1996. Penrose's Godelian argument. *Psyche* 2:21-32.

Gaifman, H. 2000. What Godel's incompleteness result does and does not show. *Journal of Philosophy* 97:462-471.

George, F. 1962. Minds, machines and Godel: Another reply to Mr. Lucas. *Philosophy* 37:62-63.

Lucas's argument applies only to deductive machines, not inductive ones.

George, A. & Velleman, D.J. 2000. Leveling the playing field between mind and machine: A reply to McCall. *Journal of Philosophy* 97:456-452.

Good, I.J. 1967. Human and machine logic. *British Journal for the Philosophy of Science* 18:145-6.

Even humans can't Godelize forever. On ordinals and transfinite counting.

Good, I.J. 1969. Godel's theorem is a red herring. *British Journal for the Philosophy of Science* 19:357-8.

Rejoinder to Lucas 1967: the role of consistency; non-constructible ordinals.

Grush, R. & Churchland, P. 1995. Gaps in Penrose's toiling. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Hanson, W. 1971. Mechanism and Godel's theorem. *British Journal for the Philosophy of Science* 22:9-16.

An analysis of Benacerraf 1967. Benacerraf's "paradox" is illusory; there are no strong consequences of Godel's theorem for mechanism.

Hofstadter, D.R. 1979. *_Godel, Escher, Bach: An Eternal Golden Braid_*. Basic Books.

Contra Lucas: we can't Godelize forever; and we're not formal on top level.

Hutton, A. 1976. This Godel is killing me. *Philosophia* 3:135-44.

Gives a statistical argument to the effect that we cannot know that we are consistent; so the Lucas argument cannot go through.

Irvine, A.D. 1983. Lucas, Lewis, and mechanism -- one more time. *Analysis* 43:94-98.

Contra Lewis 1979, Lucas can derive the consistency of M even without the premise that he is M. Hmm.

Hadley, R.F. 1987. Godel, Lucas, and mechanical models of mind. *Computational Intelligence* 3:57-63.

A nice analysis of Lucas's argument and the circumstances under which a machine might prove another's Godel sentences. There's no reason to believe that machines and humans are different here.

Jacquette, D. 1987. Metamathematical criteria for minds and machines. *Erkenntnis* 27:1-16.

A machine will fail a Turing test if it's asked about Godel sentences.

King, D. 1996. Is the human mind a Turing machine? *Synthese* 108:379-89.

Kirk, R. 1986. Mental machinery and Godel. *Synthese*.

Lucas's argument fails, as theorems by humans don't correspond to outputs of their formal systems.

Lewis, D. 1969. Lucas against mechanism. *Philosophy* 44:231-3.

Lucas needs a rule of inference from sentences to their consistency, yielding Lucas arithmetic. No machine can prove all of Lucas arithmetic, but there's no reason to suppose humans can either, as the rule is infinitary.

Lewis, D. 1979. Lucas against mechanism II. Canadian Journal of Philosophy 9:373-6.

Reply to Lucas 1970: the dialectical argument fails, as the human's output depends on the premise that it is the machine (to derive M's consistency). With a similar premise, the machine itself can do equally well.

Lucas, J.R. 1961. Minds, machines and Godel. Philosophy 36:112-127.

Humans can Godelize any given machine, so we're not a machine.

Lucas, J.R. 1967. Human and machine logic: a rejoinder. British Journal for the Philosophy of Science 19:155-6.

Reply to Good 1967: a human can trump any given machine, so the human is not the machine, whether or not the human is superior across the board.

Lucas, J.R. 1968. Satan stultified: A rejoinder to Paul Benacerraf. Monist 52:145-58.

Benacerraf 1967 is empty and omega-inconsistent. Reply to arguments based on difficulty of seeing consistency (e.g. Putnam). Fallacious but engaging.

Lucas, J.R. 1971. Metamathematics and the philosophy of mind: A rejoinder. Philosophy of Science 38:310-13.

Lucas, J.R. 1970. Mechanism: A rejoinder. Philosophy 45:149-51.

Response to Lewis 1969 and Coder 1969. Lewis misses the dialectical nature of the argument.

Lucas, J.R. 1970. The Freedom of the Will. Oxford University Press.

Lucas, J.R. 1976. This Godel is killing me: A rejoinder. Philosophia 6:145-8.

Contra Hutton, we know -- even if fallibly -- that we are consistent.

Lucas, J.R. 1984. Lucas against mechanism II: A rejoinder. Canadian Journal of Philosophy 14:189-91.

Reply to Lewis 1979.

Lucas, J.R. 1996. Mind, machines and Godel: A retrospect. In (P. Millican & A. Clark, eds) Machines and Thought. Oxford University Press.

Addresses all the counterarguments. Fun.

Lyngzeitson, A.E. & Solomon, M.K. 1994. Abstract complexity theory and the mind-machine problem. British Journal for the Philosophy of Science 45:549-54.

Lyngzeitson, A. 1990. Massively parallel distributed processing and a computationalist foundation for cognitive science. British Journal for the Philosophy of Science 41.

A Connection Machine might escape the Lucas argument. Bizarre.

Martin, J. & Engleman, K. 1990. The mind's I has two eyes. Philosophy 510-16.

Contra Hofstadter: Lucas can believe his Whitely sentence.

Maudlin, T. 1996. Between the motion and the act... Psyche 2:40-51.

McCall, S. 1999. Can a Turing machine know that the Godel sentence is true? Journal of Philosophy 96:525-32.

McCullough, D. 1996. Can humans escape Godel? Psyche 2:57-65.

McDermott, D. 1996. [Star] Penrose is wrong. Psyche 2:66-82.

Penrose, R. 1989. The Emperor's New Mind. Oxford University Press.

We are non-algorithmic as we can see Godel sentences of any algorithm.

Penrose, R. 1990. Precis of The Emperor's New Mind. Behavioral and Brain Sciences 13:643-705.

Much debate over the "non-algorithmic insight" in seeing Godel sentences.

Penrose, R. 1992. Setting the scene: The claim and the issues. In (D. Broadbent, ed) *The Simulation of Human Intelligence*. Blackwell.
An argument from the halting problem to the nonalgorithmicity of mathematical thought. Addresses objections: that the algorithm is unknowable, unsound, everchanging, environmental, or random. New physical laws may be involved.

Penrose, R. 1994. *Shadows of the Mind*. Oxford University Press.

Penrose, R. 1996. Beyond the doubting of a shadow. *Psyche* 2:89-129.
A reply to Chalmers, Feferman, Maudlin, McDermott, etc.

Priest, G. 1994. Godel's theorem and the mind... again. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind: The Place of Philosophy in the Study of Mind*. Kluwer.

Putnam, H. 1985. Reflexive reflections. *Erkenntnis* 22:143-153.
A generalized Godelian argument: if our prescriptive inductive competence is formalizable, then we could not know that such a formalization is correct.

Robinson, W.S. 1992. Penrose and mathematical ability. *Analysis* 52:80-88.
Penrose's argument depends on our knowledge of the validity of the algorithm we use, and here he equivocates between conscious and unconscious algorithms.

Slezak, P. 1982. Godel's theorem and the mind. *British Journal for the Philosophy of Science* 33:41-52.
General analysis; Lucas commits type/token error; self-ref paradoxes.

Slezak, P. 1983. Descartes's diagonal deduction. *British Journal for the Philosophy of Science* 34:13-36.
Cogito was a diagonal argument; connection to Godel, Lucas, Minsky, Nagel.

Smart, J.J.C. 1961. Godel's theorem, Church's theorem, and mechanism. *Synthese* 13:105-10.
A machine could escape the Godelian argument by inductively ascertaining its own syntax. With comments on the relevance of ingenuity.

Tymoczko, T. 1991. Why I am not a Turing Machine: Godel's theorem and the philosophy of mind. In (J. Garfield, ed) *Foundations of Cognitive Science*. Paragon House.
Weak defense of Lucas; response to Putnam, Bowie, Dennett.

Wang, H. 1974. *From Mathematics to Philosophy*. London.

Webb, J. 1968. Metamathematics and the philosophy of mind. *Philosophy of Science* 35:156-78.

Webb, J. 1980. *Mechanism, Mentalism and Metamathematics*. Kluwer.

Whitely, C. 1962. Minds, machines and Godel: A reply to Mr. Lucas. *Philosophy* 37:61-62.
Humans get trapped too: "Lucas cannot consistently assert this formula".

Yu, Q. 1992. Consistency, mechanicalness, and the logic of the mind. *Synthese* 90:145-79.

4.1c The Chinese Room (Searle)

Searle, J.R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-57.
Implementing a program is not sufficient for mentality, as someone could e.g.

implement a "Chinese-speaking" program without understanding Chinese. So strong AI is false, and no program is sufficient for consciousness.

Searle, J.R. 1984. *Minds, Brains and Science*. Harvard University Press.
Axiomatizes the argument: Syntax isn't sufficient for semantics, programs are syntactic, minds are semantic, so no program is sufficient for mind.

Searle, J.R. 1987. Minds and brains without programs. In (C. Blakemore, ed) *Mindwaves*. Blackwell.
More on the arguments against AI, e.g. the Chinese room and considerations about syntax and semantics. Mind is a high-level physical property of brain.

Searle, J.R. 1990. Is the brain's mind a computer program? *Scientific American* 262(1):26-31.
On the status of the Chinese Room argument, ten years on.

Anderson, D. 1987. Is the Chinese room the real thing? *Philosophy* 62:389-93.

Boden, M. 1988. Escaping from the Chinese Room. In *Computer Models of Mind*. Cambridge University Press.
A procedural account of how computers might have understanding and semantics.

Ben-Yami, H. 1993. A note on the Chinese room. *Synthese* 95:169-72.
A fully functional Chinese room is impossible, as it (for instance) could not say what the time is.

Bynum, T.W. 1985. Artificial intelligence, biology, and intentional states. *Metaphilosophy* 16:355-77.
A chess-playing machine embodied as a robot could have intentional states. Reference requires input/output, computation, and context.

Cam, P. 1990. Searle on strong AI. *Australasian Journal of Philosophy* 68:103-8.
Criticizes Searle's "conclusion" that brains are needed for intentionality, notes that even a homunculus has intentional states. A misinterpretation.

Carleton, L. 1984. Programs, language understanding, and Searle. *Synthese* 59:219-30.
Arguing against Searle on a number of fronts, somewhat unconvincingly.

Chalmers, D.J. 1992. Subsymbolic computation and the Chinese Room. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.
Gives an account of symbolic vs. subsymbolic computation, and argues that the latter is less vulnerable to the Chinese-room intuition, as representations there are not computational tokens.

Churchland, P.M. & Churchland, P.S. 1990. Could a machine think? *Scientific American* 262(1):32-37.
Artificial mentality is possible, not through classical AI but through brain-like AI. Argues the syntax/semantics point using an analogy with electromagnetism and luminance.

Cohen, L.J. 1986. What sorts of machines can understand the symbols they use? *Aristotelian Society Supplement* 60:81-96.

Cole, D.J. 1984. Thought and thought experiments. *Philosophical Studies* 45:431-44.
Lots of thought experiments like Searle's, against Searle. Searle's argument is like Leibniz's "mill" argument, with similar level confusions. Nice but patchy.

Cole, D.J. 1991. Artificial intelligence and personal identity. *Synthese*

88:399-417.

In the Chinese room, neither the person nor the system understands: a virtual person does. This person isn't the system, just as a normal person isn't a body. Follows from the "Kornese" room, which has two distinct understanders.

Cole, D.J. 1991. Artificial minds: Cam on Searle. Australasian Journal of Philosophy 69:329-33.

Cole, D.J. 1994. The causal powers of CPUs. In (E. Dietrich, ed) *_Thinking Computers and Virtual Persons_*. Academic Press.

Copeland, B.J. 1993. The curious case of the Chinese gym. Synthese 95:173-86. Advocates the systems reply, and criticizes Searle's "Chinese Gym" response to connectionism: Searle (like those he accuses) confuses a simulation with the thing being simulated. Nice.

Dennett, D.C. 1987. Fast thinking. In *_The Intentional Stance_*. MIT Press. Argues with Searle on many points. A little weak.

Double, R. 1983. Searle, programs and functionalism. Nature and System 5:107-14.

The homunculus doesn't have access to the system's intentionality. The syntax/semantics relation is like the neurophysiology/mind relation.

Dyer, M. 1990. Intentionality and computationalism: minds, machines, Searle and Harnad. Journal of Experimental and Theoretical Artificial Intelligence 2:303-19.

Reply to Searle/Harnad: systems reply, level confusions, etc.

Dyer, M. 1990. Finding lost minds. Journal of Experimental and Theoretical Artificial Intelligence 2:329-39.

Reply to Harnad 1990: symbols, other minds, physically embodied algorithms.

Fields, C. 1984. Double on Searle's Chinese Room. Nature and System 6:51-54. Double's argument implies that the brain isn't the basis of intentionality.

Fisher, J. 1988. The wrong stuff: Chinese rooms and the nature of understanding. Philosophical Investigations 11:279-99.

Fodor, J.A. 1991. Yin and Yang in the Chinese Room. In (D. Rosenthal, ed) *_The Nature of Mind_*. Oxford University Press.

The Chinese room isn't even implementing a Turing machine, because it doesn't use proximal causation. With a reply by Searle.

Globus, G. 1991. Deconstructing the Chinese room. Journal of Mind and Behavior 12:377-91.

Gozzano, S. 1995. Consciousness and understanding in the Chinese room. Informatica 19:653-56.

Hanna, P. 1985. Causal powers and cognition. Mind 94:53-63.

Argues that Searle is confused, and underestimates computers. Weak.

Harnad, S. 1989. Minds, machines and Searle. Journal of Experimental and Theoretical Artificial Intelligence 1:5-25.

Non-symbolic function is necessary for mentality. Trying hard to work out a theory of why the Chinese Room shows what it does. Nice but wrong.

Harnad, S. 1990. Lost in the hermeneutical hall of mirrors. Journal of Experimental and Theoretical Artificial Intelligence 2:321-27.

Reply to Dyer 1990: on the differences between real and as-if intentionality.

Hauser, L. 1997. Searle's Chinese box: Debunking the Chinese room argument.

Minds and Machines 7:199-226.

Hayes, P., Harnad, S., Perlis, D. & Block, N. 1992. Virtual symposium on virtual mind. Minds and Machines 2.

A discussion about the Chinese room, symbol grounding, and so on.

Hofstadter, D.R. 1981. Reflections on Searle. In (D. Hofstadter & D. Dennett, eds) *The Mind's I*, pp. 373-382. Basic Books.

Searle is committing a level confusion, and understates the complexity of the case. We can move from the CR to a brain (with a demon) by twiddling knobs, and the systems reply should work equally well in both cases.

Jacquette, D. 1989. Searle's intentionality thesis. *Synthese* 80:267-75.

Searle's view implies that intentional causation is not efficient causation.

Jacquette, D. 1989. Adventures in the Chinese Room. *Philosophy and Phenomenological Research* 49:605-23.

If we had microfunctional correspondence, the CR argument would fail. With points about the status of abstract/biological intentionality. A bit weak.

Searle, J.R. 1989. Reply to Jacquette. *Philosophy and Phenomenological Research* 49:701-8.

Jacquette misses the point of the argument. Also, biological and abstract intentionality are quite compatible.

Jacquette, D. 1990. Fear and loathing (and other intentional states) in Searle's Chinese Room. *Philosophical Psychology* 3:287-304.

Reply to Searle on CR, central control, biological intentionality & dualism.

Jahren, N. 1990. Can semantics be syntactic? *Synthese* 82:309-28.

Against Rapaport's Korean Room argument -- syntax isn't enough.

Korb, K. 1991. Searle's AI program. *Journal of Experimental and Theoretical Artificial Intelligence* 3:283-96.

The Chinese room doesn't succeed as an argument about semantics. At best it might succeed as an argument about consciousness.

Maloney, J.C. 1987. The right stuff. *Synthese* 70:349-72.

Defends Searle against all kinds of objections.

Melnyk, A. 1996. Searle's abstract argument against strong AI. *Synthese* 108:391-419.

Moor, J.H. 1988. The pseudorealization fallacy and the Chinese Room argument. In (J. Fetzer, ed) *Aspects of AI*. D. Reidel.

Computational systems must also meet performance criteria.

Newton, N. 1989. Machine understanding and the Chinese Room. *Philosophical Psychology* 2:207-15.

A program can possess intentionality, even if not consciousness.

Obermeier, K.K. 1983. Wittgenstein on language and artificial intelligence: The Chinese-room thought-experiment revisited. *Synthese* 56:339-50.

Pfeifer, K. 1992. Searle, strong AI, and two ways of sorting cucumbers. *Journal of Philosophical Research* 17:347-50.

Rapaport, W. 1984. Searle's experiments with thought. *Philosophy of Science* 53:271-9.

Comments on Cole, and some general material on syntax and semantics.

Rey, G. 1986. What's really going on in Searle's 'Chinese Room'. *Philosophical Studies* 50:169-85.

Recommends the systems reply, and a causal account of semantics. Discusses the relevance of wide and narrow notions of content, and the tension between Searle's positive and negative proposals.

Roberts, L. 1990. Searle's extension of the Chinese Room to connectionist machines. *Journal of Experimental and Theoretical Artificial Intelligence* 2:185-7.

In arguing against the relevance of the serial/parallel distinction to mental states, Searle becomes a formalist. A nice point.

Russow, L-M. 1984. Unlocking the Chinese Room. *Nature and System* 6:221-8. Searle's presence in the room destroys the integrity of the system, so that it is no longer a proper implementation of the program.

Seidel, A. 1988. Searle on the biological basis of cognition. *Analysis* 48:26-28.

Seidel, A. 1989. Chinese Rooms A, B and C. *Pacific Philosophical Quarterly* 20:167-73.

A person running the program, with interpretations at hand, would understand. Point-missing.

Sharvy, R. 1985. Searle on programs and intentionality. *Canadian Journal of Philosophy Supplement* 11:39-54.

Argues against Searle, but misses the point for the most part.

Sloman, A. 1986. Did Searle attack Strong Strong AI or Weak Strong AI? In (Cohn & Thomas, eds) *_Artificial Intelligence and its Applications_*. Chichester.

Suits, D. 1989. Out of the Chinese Room. *Computing and Philosophy Newsletter* 4:1-7.

Story about homunculi within homunculi. Fun.

Teng, N.Y. 2000. A cognitive analysis of the Chinese room argument. *Philosophical Psychology* 13:313-24.

Thagard, P. 1986. The emergence of meaning: An escape from Searle's Chinese Room. *Behaviorism* 14:139-46.

Get semantics computationally via induction and functional roles.

Weiss, T. 1990. Closing the Chinese room. *Ratio* 3:165-81.

Searle-in-the-room isn't in a position to know about the system's first-person states. Intrinsic intentionality is an incoherent notion.

Whitmer, J.M. 1983. Intentionality, artificial intelligence, and the causal powers of the brain. *Auslegung* 10:194-210.

Defending Searle's position, with remarks on the "causal powers" argument.

4.1d Machine Consciousness, Misc [see also 1.8b]

Angel, L. 1989. *_How to Build a Conscious Machine_*. Westview Press.

Angel, L. 1994. Am I a computer? In (E. Dietrich, ed) *_Thinking Computers and Virtual Persons_*. Academic Press.

Arrington, R. 1999. Machines, consciousness, and thought. *Idealistic Studies* 29:231-243.

Barnes, E. 1991. The causal history of computational activity: Maudlin and Olympia. *Journal of Philosophy* 88:304-16.

Response to Maudlin 1989. True computation needs active, not passive

causation, so Maudlin's machine isn't really computing.

Birnbacher, D. 1995. Artificial consciousness. In (T. Metzinger, ed) _Conscious Experience_. Ferdinand Schoningh.

Bringsjord, S. 1992. _What Robots Can and Can't Be_. Kluwer.

Bringsjord, S. 1994. Could, how could we tell if, and should -- androids have inner lives? In (K.M. Ford, C. Glymour, & P. Hayes, eds) _Android Epistemology_. MIT Press.

Caplain, G. 1995. Is consciousness a computational property? *Informatica* 19:615-19.

Coles, L.S. 1993. Engineering machine consciousness. *AI Expert* 8:34-41.

D'Aquili, E.G. & Newberg, A.B. 1996. Consciousness and the machine. *Zygon* 31:235-52.

Danto, A. 1960. On consciousness in machines. In (S. Hook, ed) _Dimensions of Mind_. New York University Press.

Dennett, D.C. 1994. The practical requirements for making a conscious robot. *Philosophical Transactions of the Royal Society A* 349:133-46.

Dennett, D.C. 1995. Cog: Steps toward consciousness in robots. In (T. Metzinger, ed) _Conscious Experience_. Ferdinand Schoningh.

Glennan, S.S. 1995. Computationalism and the problem of other minds. *Philosophical Psychology* 8:375-88.

Gunderson, K. 1968. Robots, consciousness and programmed behaviour. *British Journal for the Philosophy of Science* 19:109-22.

Gunderson, K. 1969. Cybernetics and mind-body problems. *Inquiry* 12:406-19.

Gunderson, K. 1971. _Mentality and Machines_. Doubleday.

Hillis, D. 1998. Can a machine be conscious? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness II_. MIT Press.

Kirk, R. 1986. Sentience, causation and some robots. *Australasian Journal of Philosophy* 64:308-21.

One could model brain states with monadic states and appropriate connections.
But surely that's not intelligent -- the causation has the wrong form. Nice.

Kitamura, T., Tahara, T., & Asami, K. 2000. How can a robot have consciousness? *Advanced Robotics* 14:263-275.

Lucas, J.R. 1994. A view of one's own (conscious machines). *Philosophical Transactions of the Royal Society, Series A* 349:147-52.

Maudlin, T. 1989. Computation and consciousness. *Journal of Philosophy* 86:407-32.

Computational state is not sufficient for consciousness, as it can be instantiated by a mostly inert object. A nice thought-experiment, raising questions about the relevance of counterfactuals to consciousness.

McCarthy, J. 1996. Making robots conscious of their mental states. In (S. Muggleton, ed) _Machine Intelligence 15_. Oxford University Press.

McGinn, C. 1987. Could a machine be conscious? In (C. Blakemore & S. Greenfield, ed) _Mindwaves_. Blackwell. Reprinted in _The Problem of Consciousness_ (Blackwell, 1980).

Of course, as we are machines. But what *sort* of machines are conscious, and in virtue of what properties? Remarks on artefacts, life, functionalism, and computationalism. So far, we don't know what makes the brain conscious.

Puccetti, R. 1967. On thinking machines and feeling machines. *British Journal for the Philosophy of Science* 18:39-51.

Machines can think but can't feel, so aren't persons.

Putnam, H. 1964. Robots: machines or artificially created life? *Journal of Philosophy* 61:668-91. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Various arguments and counter-arguments re machine consciousness and civil liberties. Problems of machine consciousness are analogous to problems of human consciousness. The structural basis of the two may well be the same.

Putnam, H. 1967. The mental life of some machines. In (H. Castaneda, ed) *Intentionality, Minds and Perception*. Wayne State University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

More on TMs: explaining their psychology via preference functions.

Schlagel, R. 1999. Why not artificial consciousness or thought? *Minds and Machines* 9:3-28.

Scriven, M. 1953. The mechanical concept of mind. *Mind*.

To speak of a conscious machine is to commit a semantic mistake.

Consciousness presupposes life and non-mechanism. Later retracted.

Stubenbergl, L. 1992. What is it like to be Oscar? *Synthese* 90:1-26.

Argues that AI systems like Pollock's Oscar needn't be conscious. Blindsight tells us that complex perceptual processing can go on unconsciously.

Thompson, D. 1965. Can a machine be conscious? *British Journal for the Philosophy of Science* 16:36.

Accepting machine consciousness would have few philosophical consequences, whereas rejecting it would tend to commit one to epiphenomenalism.

van de Vete, D. 1971. The problem of robot consciousness. *Philosophy and Phenomenological Research* 32:149-65.

Ziff, P. 1959. The feelings of robots. *Analysis*.

Of course robots can't think: they're not alive, so this gives us good reason not to rely on behavior. With replies by J.J.C. Smart, N. Smart.

4.1e Machine Thought, Misc

Bringsjord, S. 1998. Cognition is not computation: The argument from irreversibility. *Synthese*.

Burks, A.W. 1973. Logic, computers, and men. *Proceedings and Addresses of the American Philosophical Association* 46:39-57.

Arguing that a finite deterministic automaton can perform all natural human functions. With remarks on the logical organization of computers.

Cohen, L.J. 1955. Can there be artificial minds? *Analysis* 16:36-41.

Subservience to known or knowable rules is incompatible with mentality.

Copeland, B. J. 2000. Narrow versus wide mechanism: Including a re-examination of Turing's views on the mind-machine issue. *Journal of Philosophy* 97:5-33.

Dennett, D.C. 1985. Can machines think? In *How We Know* (Shafto).

Defends the Turing Test, among other things.

Dretske, F. 1985. Machines and the mental. Proceedings and Addresses of the American Philosophical Association 59:23-33.

Dretske, F. 1993. Can intelligence be artificial? Philosophical Studies 71:201-16.

Intelligence requires not just action or thought, but the governance of action by thought, which requires a history. "Wired-up" systems lack the explanatory connection between thought and action, so are not intelligent.

Dreyfus, H.L. 1972. _What Computers Can't Do_. Harper and Row.
Computers follow rules, people don't.

Hauser, L. 1993. Why isn't my pocket calculator a thinking thing? Minds and Machines 3:3-10.

Henley, T.B. 1990. Natural problems and artificial intelligence. Behavior and Philosophy 18:43-55.

On the philosophical importance of criteria for intelligence. With remarks on Searle, the Turing test, attitudes to AI, and ethical considerations.

Kearns, J.T. 1997. Thinking machines: Some fundamental confusions. Minds and Machines 7:269-87.

Lanier, J. 1998. Three objections to the idea of artificial intelligence.
In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness II_. MIT Press.

Mackay, D.M. 1951. Mind-life behavior in artifacts. British Journal for the Philosophy of Science 2:105-21.

Mackay, D.M. 1952. Mentality in machines. Aristotelian Society Supplement 26:61-86.

Manning, R.C. 1987. Why Sherlock Holmes can't be replaced by an expert system. Philosophical Studies 51:19-28.

An expert system would lack Holmes' ability to raise the right questions, sort out relevant data, and determine what data are in need of explanation.

Mays, W. 1952. Can machines think? Philosophy 27:148-62.

McCarthy, J. 1979. Ascribing mental qualities to machines. In (M. Ringle, ed) _Philosophical Perspectives in Artificial Intelligence_. Humanities Press.

Negley, G. 1951. Cybernetics and theories of mind. Journal of Philosophy 48:574-82.

Preston, B. 1995. The ontological argument against the mind-machine hypothesis. Philosophical Studies 80:131-57.

Lucas, Searle, and Penrose all fall prey to "dual-description" fallacies.

Puccetti, R. 1966. Can humans think? Analysis.

Rapaport, W. 1993. Because mere calculating isn't thinking: Comments on Hauser's "Why isn't my pocket calculator a thinking thing?". Minds and machines 3:11-20.

Scriven, M. 1960. The compleat robot: A prolegomena to androidology. In (S. Hook, ed) _Dimensions of Mind_. New York University Press.

A machine could possess every characteristic of human thought: e.g. freedom, creativity, learning, understanding, perceiving, feeling.

Spilsbury, R.J. 1952. Mentality in machines. Aristotelian Society Supplement 26:27-60.

4.2 Computation and Representation

4.2a Symbols and Symbol Systems [see also 2.1a, 4.3e]

Cummins, R. 1996. Why there is no symbol grounding problem? In Representations, Targets, and Attitudes. MIT Press.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42:335-346.
AI symbols are empty and meaningless. They need to be "grounded" in something, e.g. sensory projection. Maybe connectionism can do the trick?

Harnad, S. 1992. Connecting object to symbol in modeling cognition. In (A. Clark & R. Lutz, eds) Connectionism in Context. Springer-Verlag.
On the limitations of symbol systems, and the potential for grounding symbols in sensory icons and categorical perception, e.g. with neural networks.

Kosslyn, S.M. & Hatfield, G. 1984. Representation without symbol systems. *Social Research* 51:1019-1045.

MacDorman, K.F. 1997. How to ground symbols adaptively. In (S. O'Neillain, P. McKevitt, & E. MacAogain, eds) Two Sciences of Mind. John Benjamins.

Newell, A. 1980. Physical symbol systems. *Cognitive Science* 4:135-83.

Newell, A. & Simon, H.A. 1981. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery* 19:113-26. Reprinted in (J. Haugeland, ed) Mind Design. MIT Press.
On computer science, AI, & the Physical Symbol System Hypothesis.

Robinson, W.S. 1995. Brain symbols and computationalist explanation. *Minds and Machines* 5:25-44.

Sun, R. 2000. Symbol grounding: a new look at an old idea. *Philosophical Psychology* 13:149-172.

4.2b Computational Semantics

Fodor, J.A. 1978. Tom Swift and his procedural grandmother. *Cognition* 6:229-47. Reprinted in RePresentations (MIT Press, 1980).
Against procedural semantics; it's a rerun of verificationism.

Hadley, R.F. 1990. Truth conditions and procedural semantics. In (P. Hanson, ed) Information, Language and Cognition. University of British Columbia Press.

Johnson-Laird, P. 1977. Procedural semantics. *Cognition* 5:189-214.

Johnson-Laird, P. 1978. What's wrong with Grandma's guide to procedural semantics: A reply to Jerry Fodor. *Cognition* 9:249-61.

McDermott, D. 1978. Tarskian semantics, or no notation without denotation. *Cognitive Science* 2:277-82.
On the virtues of denotational semantics for AI. Notation without denotation, as found in many AI systems, leads to castles in the air.

Perlis, D. 1991. Putting one's foot in one's head -- Part 1: Why. *Nous* 25:435-55.

Perlis, D. 1994. Putting one's foot in one's head -- Part 2: How. In (E.

Dietrich, ed) *_Thinking Computers and Virtual Persons_*. Academic Press.

Rapaport, W.J. 1988. Syntactic semantics: Foundations of computational natural language understanding. In (J. Fetzer, ed) *_Aspects of AI_*. Kluwer.

Rapaport, W.J. 1995. Understanding understanding: Syntactic semantics and computational cognition. *Philosophical Perspectives* 9:49-88.

Smith, B. 1988. On the semantics of clocks. In (J. Fetzer, ed) *_Aspects of AI_*. Kluwer.

Smith, B. 1987. The correspondence continuum. *CSLI-87-71*.

Wilks, Y. 1982. Some thoughts on procedural semantics. In (W. Lehnert, ed) *_Strategies for Natural Language Processing_*. Lawrence Erlbaum.

Wilks, Y. 1990. Form and content in semantics. *Synthese* 82:329-51.
Criticism of McDermott's views on semantics, logic and natural language.

Winograd, T. 1985. Moving the semantic fulcrum. *Linguistics and Philosophy* 8:91-104.

Woods, W. 1981. Procedural semantics as a theory of meaning. In (A. Joshi, B. Weber, & I. Sag) *_Elements of Discourse Understanding_*. Cambridge University Press.

Woods, W. 1986. Problems in procedural semantics. In (Z. Pylyshyn & W. Demopolous, eds) *_Meaning and Cognitive Structure_*. Ablex.
With commentaries by Haugeland, J.D. Fodor.

4.2c Implicit/Explicit Rules and Representations

Clark, A. 1991. In defense of explicit rules. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.
Argues that we need explicit rules for flexibility, adaptability, and representational redescription. With remarks on eliminativism.

Cummins, R. 1986. Inexplicit information. In (M. Brand & R. Harnish, eds) *_The Representation of Knowledge and Belief_*. University of Arizona Press.
On various kinds of representation of knowledge or belief without explicit tokens: control-implicit, domain-implicit, and procedural information.
The key distinction is representation vs. execution of a rule.

Davies, M. 1995. Two notions of implicit rules. *Philosophical Perspectives* 9:153-83.

Hadley, R.F. 1990. Connectionism, rule-following, and symbolic manipulation. *Proc AAAI*.
Some rules are learnt so quickly that representation must be explicit.

Hadley, R.F. 1993. Connectionism, explicit rules, and symbolic manipulation. *Minds and Machines* 3.

Hadley, R.F. 1995. The 'explicit-implicit' distinction. *Minds and Machines* 5:219-42.

Kirsh, D. 1990. When is information explicitly represented? In (P. Hanson, ed) *_Information, Language and Cognition_*. University of British Columbia Press.

Skokowski, P.G. 1994. Can computers carry content "inexplicitly"? *Minds and Machines* 4:333-44.
Cummins' account of inexplicit information fails, as even "executed" rules

must be represented in the system. With remarks on the Chinese room.

4.2d AI without Representation?

Bechtel, W. 1996. Yet another revolution? Defusing the dynamical system theorists' attack on mental representations. Manuscript.

Brooks, R. 1991. Intelligence without representation. Artificial Intelligence 47:139-159.

We don't need explicit representation; the world can do the job instead.
Use embodied, complete systems, starting simple and working incrementally.

Clark, A. and Toribio, J. 1994. Doing without representing. Synthese 101:401-31.

A discussion of anti-representationalism in situated robotics and the dynamic systems movement (Brooks, Beer, van Gelder). These arguments appeal to overly simple domains, and a modest notion of representation survives.

Keijzer, F.A. 1998. Doing without representations which specify what to do. Philosophical Psychology 11:269-302.

Kirsh, D. 1991. Today the earwig, tomorrow man? Artificial Intelligence 47:161-184.

van Gelder, T. 1995. What might cognition be if not computation? Journal of Philosophy 92:345-81.

Argues for a dynamic-systems conception of the mind that is non-computational and non-representational. Uses an analogy with the Watt steam governor to argue for a new kind of dynamic explanation.

4.2e Miscellaneous

Chrisley, R.L. 1994. Taking embodiment seriously: Nonconceptual content and robotics. In (K.M. Ford, C. Glymour, & P. Hayes, eds) Android Epistemology. MIT Press.

Dietrich, E. 1988. Computers, intentionality, and the new dualism. Manuscript.

Dreyfus, H.L. 1979. A framework for misrepresenting knowledge. In (M. Ringle, ed) Philosophical Perspectives in Artificial Intelligence. Humanities Press.
On the problems with context-free symbolic representation.

Fields, C. 1994. Real machines and virtual intentionality: An experimentalist takes on the problem of representational content. In (E. Dietrich, ed) Thinking Computers and Virtual Persons. Academic Press.

Haugeland, J. 1981. Semantic engines: An introduction to mind design. In (J. Haugeland, ed) Mind Design. MIT Press.

Robinson, W.S. 1995. Direct representation. Philosophical Studies 80:305-22.
On Searle's critique of computational explanation, contrasted with Gallistel's use thereof. The real issue is computation on indirect vs. direct representations; direct computationalism is an attractive view.

4.3 Philosophy of Connectionism

4.3a Connectionism and Compositionality (Fodor/Pylyshyn)

Fodor, J.A. & Pylyshyn, Z.W. 1988. Connectionism and cognitive architecture. *Cognition* 28:3-71.

Connectionist models can't explain cognitive systematicity and productivity, as their representations lack compositional structure. The allures of connectionism are illusory; it's best used as an implementation strategy.

Aizawa, K. 1997. Explaining systematicity. *Mind and Language* 12:115-36.

Aizawa, K. 1997. The role of the systematicity argument in classicism and connectionism. In (S. O'Neill, ed) *Two Sciences of Mind*. John Benjamins.

Aizawa, K. 1997. Exhibiting versus explaining systematicity: A reply to Hadley and Hayward. *Minds and Machines* 7:39-55.

Antony, M. 1991. Fodor and Pylyshyn on connectionism. *Minds and Machines* 1:321-41.

Fodor and Pylyshyn's argument is an invalid instance of inference to the best explanation, as there is much to explain than systematicity. Connectionism and classicism may be compatible even without implementation, in any case.

Aydede, M. 1997. Language of thought: The connectionist contribution. *Minds and Machines* 7:57-101.

Butler, K. 1991. Towards a connectionist cognitive architecture. *Mind and Language* 6:252-72.

Connectionism can make do with unstructured representations, as long as they have the right causal relations between them.

Butler, K. 1993. Connectionism, classical cognitivism, and the relation between cognitive and implementational levels of analysis. *Philosophical Psychology* 6:321-33.

Contra Chalmers 1993, F&P's argument doesn't apply at the implementational level. Contra Chater and Oaksford 1990, connectionism can't be purely implementational, but some implementational details can be relevant.

Butler, K. 1993. On Clark on systematicity and connectionism. *British Journal for the Philosophy of Science* 44:37-44.

Argues against Clark on holism and the conceptual truth of systematicity.

Butler, K. 1995. Compositionality in cognitive models: The real issue. *Philosophical Studies* 78:153-62.

Chalmers, D.J. 1990. Syntactic transformations on distributed representations. *Connection Science* 2:53-62.

An experimental demonstration that connectionist models can handle structure-sensitive operations in a non-classical way, transforming structured representations of active sentences to passive sentences.

Chalmers, D.J. 1993. Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology* 6:305-319.

Points out a structural flaw in F&P's argument, and traces the problem to a lack of appreciation of distributed representation. With some empirical results on structure sensitive processing, and some remarks on explanation.

Chater, N. & Oaksford, M. 1990. Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition* 34:93-107.

Implementation can make a difference at the algorithmic level.

Christiansen, M.H. & Chater, N. 1994. Generalization and connectionist language learning. *Mind and Language* 9:273-87.

Cummins, R. 1996. Systematicity. *Journal of Philosophy* 93:591-614.

Fetzer, J.H. 1992. Connectionism and cognition: Why Fodor and Pylyshyn are wrong. In (A. Clark & R. Lutz, eds) *_Connectionism in Context_*. Springer-Verlag.

Fodor, J.A. & McLaughlin, B.P. 1990. Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition* 35:183-205.
Smolensky's weak compositionality is useless; and tensor product architecture can't support systematicity, as nonexistent tokens can't play a causal role.

Fodor, J.A. 1997. Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition* 62:109-19.

Garcia-Carpintero, M. 1996. Two spurious varieties of compositionality. *Minds and Machines* 6:159-72.

Garfield, J. 1997. Mentalese not spoken here: Computation, cognition, and causation. *Philosophical Psychology* 10:413-35.

Guarini, M. 1996. Tensor products and split-level architecture: Foundational issues in the classicism-connectionism debate. *Philosophy of Science* 63:S239-47.

Hadley, R.F. 1997. Cognition, systematicity, and nomic necessity. *Mind and Language* 12:137-53.

Hadley, R.F. 1994. Systematicity in connectionist language learning. *Mind and Language* 9:247-72.
Argues that existing connectionist models do not achieve an adequate systematicity in learning; they fail to generalize to handle structures with novel constituents.

Hadley, R.F. 1994. Systematicity revisited. *Mind and Language* 9:431-44.

Hadley, R.F. & Hayward, M.B. 1997. Strong semantic systematicity from Hebbian connectionist learning. *Minds and Machines* 7:1-55.

Hadley, R.F. 1997. Cognition, systematicity, and nomic necessity. *Mind and Language* 12:137-53.

Hadley, R.F. 1997. Explaining systematicity: A reply to Kenneth Aizawa. *Minds and Machines* 12:571-79.

Hawthorne, J. 1989. On the compatibility of connectionist and classical models. *Philosophical Psychology* 2:5-16.
Localist connectionist models may not be able to handle structured presentation, but appropriate distributed models can.

Horgan, T. & Tienson, J. 1991. Structured representations in connectionist systems? In (Davis, ed) *_Connectionism: Theory and Practice_*.
A discussion of how connectionism might achieve "effective syntax" without implementing a classical system.

Matthews, R.J. 1994. Three-concept monte: Explanation, implementation, and systematicity. *Synthese* 101:347-63.
F&P deal a sucker bet: on their terms, connectionism could never give a non-implementational explanation of systematicity, as the notions are construed in a manner specific to classical architectures.

Matthews, R.J. 1997. Can connectionists explain systematicity? *Mind and Language* 12:154-77.

McLaughlin, B.P. 1992. Systematicity, conceptual truth, and evolution. In *_Philosophy and the Cognitive Sciences_*.

Against responses to Fodor and Pylyshyn claiming that cognitive theories needn't explain systematicity. Contra Clark, the conceptual truth of systematicity won't help. Contra others, nor will evolution.

McLaughlin, B.P. 1993. The connectionism/classicism battle to win souls. *Philosophical Studies* 71.

Argues that no connectionist model so far has come close to explaining systematicity. Considers the models of Elman, Chalmers, and Smolensky.

Niklasson, L.F. & van Gelder, T. 1994. On being systematically connectionist. *Mind and Language* 9:288-302.

Pollack, J.B. 1990. Recursive distributed representations. *Artificial Intelligence* 46:77-105.

Develops a connectionist architecture -- recursive auto-associative memory -- that can recursively represent compositional structures in distributed form.

Rowlands, M. 1994. Connectionism and the language of thought. *British Journal for the Philosophy of Science* 45:485-503.

F&P's argument confuses constituent structure with logical/sentential structure. Connectionism is a psychotechnic project, whereas propositional description is a psychosemantic project.

Schroder, J. 1998. Knowledge of rules, causal systematicity, and the language of thought. *Synthese* 117:313-330.

Smolensky, P. 1987. The constituent structure of connectionist mental states. *Southern Journal of Philosophy Supplement* 26:137-60.

F&P ignore distributed representation and interaction effects.

Smolensky, P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159-216.

Develops a connectionist architecture that represents compositional structures as tensor products of distributed representations.

Smolensky, P. 1991. Connectionism, constituency and the language of thought. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*.

Blackwell.

Connectionism can do compositionality its own way, including both weak compositionality (with context effects) or strong compositionality (via tensor products).

Smolensky, P. 1995. Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In (C. Macdonald, ed)

Connectionism: Debates on Psychological Explanation. Blackwell.

van Gelder, T. 1990. Compositionality: A connectionist variation on a classical theme. *Cognitive Science* 14:355-84.

Connectionism can do compositionality functionally. All one needs is the right functional relation between representations; physical concatenation is not necessary.

van Gelder, T. 1991. Classical questions, radical answers. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

On connectionism as a Kuhnian paradigm shift in cognitive science, with emphasis on the implications of functional compositionality and distributed representations.

4.3b Representation in Connectionism

Butler, K. 1995. Representation and computation in a deflationary assessment

of connectionist cognitive science. *Synthese* 104:71-97.

Clark, A. 1989. Connectionism, nonconceptual content, and representational redescription. Manuscript.

On some troubles connectionism has with higher-order knowledge. Contrasts Cussins, Karmiloff-Smith on development. Subsymbols without symbols are blind.

Clark, A. 1993. *_Associative Engines: Connectionism, Concepts, and Representational Change_*. MIT Press.

Clark, A. & Karmiloff-Smith, A. 1994. The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language* 8:487-519.

On the importance of representational redescription, and on the limits of connectionist networks in cross-domain knowledge transfer. What does a true believer need, above behavior: conceptual combination, real-world fluency?

Cummins, R. 1991. The role of representation in connectionist explanation of cognitive capacities. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.

Connectionism isn't really radical. There's no new concept of representation or of learning, and cognition can still be the manipulation of semantically structured representations.

Cussins, A. 1990. The connectionist construction of concepts. In (M. Boden, ed) *_The Philosophy of AI_*. Oxford University Press.

Connectionism builds up concepts from the nonconceptual level. From nonconceptual content (e.g. perceptual experiences) to the emergence of objectivity.

Garzon, F. 2000. A connectionist defence of the inscrutability thesis. *Mind and Language* 15:465-480.

Garzón, F. 2000. State space semantics and conceptual similarity: reply to Churchland. *Philosophical Psychology* 13:77-96.

Goschke, T. & Koppelberg, D. 1990. Connectionism and the semantic content of internal representation. *Review of International Philosophy* 44:87-103.

Goschke, T. & Koppelberg, D. 1991. The concept of representation and the representation of concepts in connectionist models. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.

On correlational semantics and context-dependent representations.

Hatfield, G. 1991. Representation and rule-instantiation in connectionist systems. In (T. Horgan & J. Tienson, eds) *_Connectionism and the Philosophy of Mind_*. Kluwer.

Some remarks on psychology & physiology. Even connectionism uses psychological concepts.

Hatfield, G. 1991. Representation in perception and cognition: Connectionist affordances. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.

Haybron, D.M. 2000. The causal and explanatory role of information stored in connectionist networks. *Minds and Machines* 10:361-380.

Laakso, A. & Cottrell, G. 2000. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical Psychology* 13:47-76.

Place, U.T. 1989. Toward a connectionist version of the causal theory of

reference. Acta Analytica 4:71-97.

Ramsey, W. 1995. Rethinking distributed representation. Acta Analytica 10:9-25.

Ramsey, W. 1997. Do connectionist representations earn their explanatory keep? Mind and Language 12:34-66.

Argues that talk of representations has no explanatory role in connectionist theory, and can be discarded. It can't be understood along the lines of the teleo-informational or classical frameworks.

Schopman, J. & Shawky, A. 1996. Remarks on the impact of connectionism on our thinking about concepts. In (P. Millican & A. Clark, eds) _Machines and Thought_. Oxford University Press.

Tye, M. 1987. Representation in pictorialism and connectionism. Southern Journal of Philosophy Supplement 26:163-184.

Pictorialism isn't compatible with language of thought, but connectionism might be.

van Gelder, T. 1991. What is the D in PDP? In (W. Ramsey, S. Stich, & D. Rumelhart, eds) _Philosophy and Connectionist Theory_. Lawrence Erlbaum.

Argues that distributed representation is best analyzed in terms of superposition of representation, not in terms of extendedness.

4.3c Connectionism and Eliminativism

Ramsey, W., Stich, S.P. & Garon, J. 1991. Connectionism, eliminativism and the future of folk psychology. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) _Philosophy and Connectionist Theory_. Lawrence Erlbaum.

Connectionism implies eliminativism, as connectionist systems do not have functionally discrete contentful states, and folk psychology is committed to functional discreteness of propositional attitudes.

Bickle, J. 1993. Connectionism, eliminativism, and the semantic view of theories. Erkenntnis.

Outlines the semantic view of scientific theories, and applies it to the connectionism/eliminativism debate. There's no reason why folk psychology shouldn't be reducible, in a homogeneous or heterogeneous way.

Botterill, G. 1994. Beliefs, functionally discrete states, and connectionist networks. British Journal for the Philosophy of Science 45:899-906.

Distinguishes active from dispositional beliefs: the former are realized discretely in activation patterns, the latter nondiscretely in weights, which is all that folk psychology needs.

Clapin, H. 1991. Connectionism isn't magic. Minds and Machines 1:167-84.

Commentary on Ramsey/Stich/Garon. Connectionism has symbols that interact, and has propositional modularity in processing if not in storage.

Clark, A. 1989. Beyond eliminativism. Mind and Language 4:251-79.

Connectionism needn't imply eliminativism, as higher levels may have a causal role, if not causal completeness. Also, it may not tell the whole story.

Clark, A. 1990. Connectionist minds. Proceedings of the Aristotelian Society 90:83-102.

Responding to eliminativist challenge via cluster analysis and recurrence.

Davies, M. 1991. Concepts, connectionism, and the language of thought.

(W. Ramsey, S. Stich, & D. Rumelhart, eds) _Philosophy and Connectionist Theory_. Lawrence Erlbaum.

Argues that our conception of thought requires causal systematicity, which

requires a language of thought. Connectionist systems are not causally systematic, so connectionism leads to eliminativism.

Egan, F. 1995. Folk psychology and cognitive architecture. *Philosophy of Science* 62:179-96.

Forster, M. & Saidel, E. 1994. Connectionism and the fate of folk psychology. *Philosophical Psychology* 7:437-52.

Contra Ramsey, Stich, and Garon, connectionist representations can be seen to be functionally discrete on an appropriate analysis of causal relevance.

Horgan, T., and Tienson, J. 1995. Connectionism and the commitments of folk psychology. *Philosophical Perspectives* 9:127-52.

O'Brien, G. 1991. Is connectionism commonsense? *Philosophical Psychology* 4:165-78.

O'Leary-Hawthorne, J. 1994. On the threat of eliminativism. *Philosophical Studies* 74:325-46.

A dispositional construal of beliefs and desires can distinguish the relevant active states (via counterfactuals) and is compatible with FP, so internals can't threaten FP. With remarks on Davidson, overdetermination, etc.

Place, U.T. 1992. Eliminative connectionism: Its implications for a return to an empiricist/behaviorist linguistics. *Behavior and Philosophy* 20:21-35.

Ramsey, W. 1994. Distributed representation and causal modularity: A rejoinder to Forster and Saidel. *Philosophical Psychology* 7:453-61.

Upon examination, the model of Forster and Saidel 1994 does not exhibit features that are both distributed and causally discrete.

Smolensky, P. 1995. On the projectable predicates of connectionist psychology: A case for belief. In (C. Macdonald, ed) *_Connectionism: Debates on Psychological Explanation_*. Blackwell.

Stich, S. & Warfield, T. 1995. Reply to Clark and Smolensky: Do connectionist minds have beliefs? In (C. Macdonald, ed) *_Connectionism: Debates on Psychological Explanation_*. Blackwell.

4.3d The Connectionist/Classical Debate

Adams, F., Aizawa, K. & Fuller, G. 1992. Rules in programming languages and networks. In (J. Dinsmore, ed) *_The Symbolic and Connectionist Paradigms: Closing the Gap_*. Lawrence Erlbaum.

The distinction between programming languages and networks is neutral on rule-following, etc, so there's nothing really new about connectionism.

Aizawa, K. 1994. Representations without rules, connectionism, and the syntactic argument. *Synthese* 101:465-92.

Bringsjord, S. 1991. Is the connectionist-logician debate one of AI's wonderful red herrings? *Journal of Theoretical and Experimental Artificial Intelligence* 3:319-49.

A detailed analysis purporting to show that connectionism and "logicism" are compatible, as Turing machines can do everything a neural network can. Entertaining, but misunderstands subsymbolic processing.

Broadbent, D. 1985. A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General* 114:189-92.

Distributed models are at the implementational, not computational, level.

Chandrasekaran, B., Goel, A. & Allemang, D. 1988. Connectionism and

information-processing abstractions. AI Magazine 24-34.

Connectionism won't affect AI too much, as AI is concerned with the information-processing (task) level. With greater modularity, connectionism will look more like traditional AI.

Corbi, J.E. 1993. Classical and connectionist models: Levels of description. Synthese 95:141-68.

Dawson, M.R.W., Medler, D.A., & Berkeley, I.S.N. 1997. PDP networks can provide models that are not mere implementations of classical theories. Philosophical Psychology 10:25-40.

Dennett, D.C. 1986. The logical geography of computational approaches: A view from the east pole. In (M. Brand & R. Harnish, eds) _The Representation of Knowledge and Belief_. University of Arizona Press.

Drawing the battle-lines: High Church Computationalism at the "East Pole", New Connectionism, Zen Holism, etc, at various locations on the "West Coast". With remarks on connectionism, and on AI as thought-experimentation.

Dennett, D.C. 1991. Mother Nature versus the walking encyclopedia. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) _Philosophy and Connectionist Theory_. Lawrence Erlbaum.

Reiterating the value of connectionism, especially biological plausibility.

Dinsmore, J. (ed) 1992. _The Symbolic and Connectionist Paradigms: Closing the Gap_. Lawrence Erlbaum.

Dyer, M. 1991. Connectionism versus symbolism in high-level cognition. In (T. Horgan & J. Tienson, eds) _Connectionism and the Philosophy of Mind_. Kluwer.

Garson, J.W. 1991. What connectionists cannot do: The threat to Classical AI. In (T. Horgan & J. Tienson, eds) _Connectionism and the Philosophy of Mind_. Kluwer.

Connectionism and classicism aren't necessarily incompatible on symbolic discreteness, causal role, functional discreteness, constituency, representation of rules.

Garson, J.W. 1994. No representations without rules: The prospects for a compromise between paradigms in cognitive science. Mind and Language 9:25-37.

Garson, J.W. 1994. Cognition without classical architecture. Synthese 100:291-306.

Horgan, T. & Tienson, J. 1987. Settling into a new paradigm. Southern Journal of Philosophy Supplement 26:97-113.

On connectionism, basketball, and representation without rules. Responses to the "syntactic" and "semantic" arguments against connectionism. Nice.

Horgan, T. & Tienson, J. 1989. Representation without rules. Philosophical Perspectives 17:147-74.

Cognition uses structured representations without high-level rules, and connectionism is better at accounting for this. With remarks on exceptions to psychological laws, and the crisis in traditional AI.

Horgan, T. & Tienson, J. 1994. Representations don't need rules: Reply to James Garson. Mind and Language 9:1-24.

McClelland, J.L. & Rumelhart, D.E. 1985. Levels indeed! A response to Broadbent. Journal of Experimental Psychology: General 114:193-7.

Response to Broadbent 1985: Distributed models are at the algorithmic level. Elucidating the low-level/high-level relation via various analogies.

McLaughlin, B.P. & Warfield, F. 1994. The allure of connectionism reexamined. *Synthese* 101:365-400.

Argues that symbolic systems such as decision trees are as good at learning and pattern recognition as connectionist networks, and it is just as plausible that they are implemented in the brain.

Rey, G. 1991. An explanatory budget for connectionism and eliminativism. In (T. Horgan & J. Tienson, eds) *_Connectionism and the Philosophy of Mind_*. Kluwer.

Challenges connectionism to explain things that the classical approach seems to handle better: the structure, systematicity, causal role, and grain of propositional attitudes, their rational relations, and conceptual stability.

4.3e Subsymbolic Computation (Smolensky)

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1-23.

Connectionism offers a complete account at the subsymbolic level, rather than an approximate account at the symbolic level.

Berkeley, I. 2000. What the #\$*%! is a subsymbol? *Minds and machines* 10:1-14.

Chalmers, D.J. 1992. Subsymbolic computation and the Chinese Room. In (J. Dinsmore, ed) *_The Symbolic and Connectionist Paradigms: Closing the Gap_*. Lawrence Erlbaum.

Explicates the distinction between symbolic and subsymbolic computation, and argues that connectionism can better handle the emergence of semantics from syntax, due to the non-atomic nature of its representations.

Clark, A. 1993. Superpositional connectionism: A reply to Marinov. *Minds and Machines* 3:271-81.

Hofstadter, D.R. 1983. Artificial intelligence: Subcognition as computation. In (F. Machlup, ed) *_The Study of Information: Interdisciplinary Messages_*. Wiley. Reprinted as "Waking up from the Boolean dream" in *_Metamagical Themas_*. Basic Books.

AI needs statistical emergence. For real semantics, symbols must be decomposable, complex, autonomous -- i.e. active.

Marinov, M. 1993. On the spuriousness of the symbolic/subsymbolic distinction. *Minds and Machines* 3:253-70.

Argues with Smolensky: symbolic systems such as decision trees have all the positive features of neural networks (flexibility, lack of brittleness), and can represent concepts as sets of subconcepts. With a reply by Clark.

Rosenberg, J. 1990. Treating connectionism properly: Reflections on Smolensky. *Psychological Research* 52:163.

Rejects Smolensky's PTC, as the proper interaction of the microscopic and macroscopic levels would take a "miracle".

Smolensky, P. 1987. Connectionist AI, symbolic AI, and the brain. *AI Review* 1:95-109.

On connectionist networks as subsymbolic dynamic systems.

4.3f Philosophy of Connectionism, Misc.

Bechtel, W. 1985. Are the new PDP models of cognition cognitivist or associationist? *Behaviorism* 13:53-61.

Bechtel, W. 1986. What happens to accounts of mind-brain relations if we forgo an architecture of rules and representations? *Philosophy of Science*

Association 1986, 159-71.

On the relationship between connectionism, symbol processing, psychology and neuroscience.

Bechtel, W. 1987. Connectionism and the philosophy of mind. Southern Journal of Philosophy Supplement 26:17-41. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Lots of questions about connectionism.

Bechtel, W. 1988. Connectionism and rules and representation systems: Are they compatible? *Philosophical Psychology* 1:5-16.

There's room for both styles within a single mind. The rule-based level needn't be autonomous; the connectionist level plays a role in pattern recognition, concepts, and so on.

Bechtel, W. & Abrahamson, A. 1990. Beyond the exclusively propositional era. *Synthese* 82:223-53.

An account of the shift from propositions to pattern recognition in the study of cognition: knowing-how, imagery, categorization, connectionism.

Bechtel, W. & Abrahamsen, A.A. 1992. Connectionism and the future of folk psychology. In (R. Burton, ed) *Minds: Natural and Artificial*. SUNY Press.

Bechtel, W. 1993. The case for connectionism. *Philosophical Studies* 71:119-54.

Bickle, J. 1995. Connectionism, reduction, and multiple realizability. *Behavior and Philosophy* 23:29-39.

Bradshaw, D.E. 1991. Connectionism and the specter of representationalism. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Argues that connectionism allows for a more plausible epistemology of perception, compatible with direct realism rather than representationalism. With remarks on Fodor and Pylyshyn's argument against Gibson.

Churchland, P.M. 1989. On the nature of theories: A neurocomputational perspective. *Minnesota Studies in the Philosophy of Science* 14. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Connectionism will revolutionize our review of scientific theories: >From the deductive-nomological view to descent in weight-space. Some cute analogies.

Churchland, P.M. 1989. On the nature of explanation: A PDP approach. In *A Neurocomputational Perspective*. MIT Press.

We achieve explanatory understanding not through the manipulation of propositions but through the activation of prototypes.

Churchland, P.S. & Sejnowski, T. 1989. Neural representation and neural computation. In (L. Nadel, ed) *Neural Connections, Mental Computations*. MIT Press.

Implications of connectionism and neuroscience for our concept of mind.

Clark, A. 1989. *Microcognition*. MIT Press.

All kinds of stuff on connectionism and philosophy.

Clark, A. 1990. Connectionism, competence and explanation. *British Journal for the Philosophy of Science* 41:195-222.

Connectionism separates processing from competence. Instead of hopping down Marr's levels (theory->process), connectionism goes (1) task (2) low-level performance (3) extract theory from process. Cute.

Cummins, R. & Schwarz, G. 1987. Radical connectionism. *Southern Journal of Philosophy Supplement* 26:43-61.

On computation and representation in AI and connectionism, and on problems for radical connectionism in reconciling these without denying representation or embracing mystery.

Cummins, R. & Schwarz, G. 1991. Connectionism, computation, and cognition. In (T. Horgan & J. Tienson, eds) *_Connectionism and the Philosophy of Mind_*. Kluwer.

Explicates computationalism, and discusses ways in which connectionism might end up non-computational: if causal states cross-classify representational states, or if transitions between representations aren't computable.

Cummins, R. 1995. Connectionist and the rationale constraint on cognitive explanations. *Philosophical Perspectives* 9:105-25.

Davies, M. 1989. Connectionism, modularity and tacit knowledge. *British Journal for the Philosophy of Science* 40:541-55.

Argues that connectionist networks don't have tacit knowledge of modular theories (as representations lack the appropriate structure, etc.).

Globus, G.G. 1992. Derrida and connectionism: Differance in neural nets. *Philosophical Psychology* 5:183-97.

Hatfield, G. 1990. Gibsonian representations and connectionist symbol-processing: prospects for unification. *Psychological Research* 52:243-52.

Gibson is compatible with connectionism. In both, we can have rule-instantiation without rule-following.

Horgan, T. & Tienson, J. (eds) 1991. *_Connectionism and the Philosophy of Mind_*. Kluwer.

Horgan, T. & Tienson, J. 1996. *_Connectionism and the Philosophy of Psychology_*. MIT Press.

Horgan, T. 1997. Connectionism and the philosophical foundations of cognitive science. *Metaphilosophy* 28:1-30.

Humphreys, G.W. 1986. Information-processing systems which embody computational rules: The connectionist approach. *Mind and Language* 1:201-12.

Legg, C.R. 1988. Connectionism and physiological psychology: A marriage made in heaven? *Philosophical Psychology* 1:263-78.

Litch, M. 1997. Computation, connectionism and modelling the mind. *Philosophical Psychology* 10:357-364.

Lloyd, D. 1989. Parallel distributed processing and cognition: Only connect? In *_Simple Minds_*. MIT Press.

An overview: local/distributed/featural representations; explanation in connectionism (how to avoid big mush); relation to neuroscience; explicit representations of rules vs weight matrices.

Lycan, W.G. 1991. Homuncular functionalism meets PDP. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.

On various ways in which connectionism relates to representational homuncular functionalism, e.g. on implementation, eliminativism, and explanation.

Macdonald, C. 1995. *_Connectionism: Debates on Psychological Explanation_*. Blackwell.

Ramsey, W. & Stich, S.P. 1990. Connectionism and three levels of nativism. *Synthese* 82:177-205.

How connectionism bears on the nativism debate. Conclusion: not too much.

Ramsey, W., Stich, S.P. & Rumelhart, D.M. (eds) 1991. *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Rosenberg, J. 1989. Connectionism and cognition. Bielefeld Report.
Criticism of Churchland's connectionist epistemology.

Sehon, S. 1998. Connectionism and the causal theory of action explanation. *Philosophical Psychology* 11:511-532.

Shanon, B. 1992. Are connectionist models cognitive? *Philosophical Psychology*.

In some senses of "cognitive", yes; in other senses, no. Phenomenological, theoretical, and sociological perspectives. Toward meaning-laden models.

Sterelny, K. 1990. Connectionism. In *The Representational Theory of Mind*. Blackwell.

Waskan, J. & Bechtel, W. 1997. Directions in connectionist research: Tractable computations without syntactically structured representations. *Metaphilosophy* 28:31-62.

4.3g Foundational Empirical Issues

Clark, A. 1994. Representational trajectories in connectionist learning. *Minds and Machines* 4:317-32.

On how to get connectionist networks to learn about structured task domains. Concentrates on incremental learning, and other developmental/scaffolding strategies. With remarks on systematicity.

Clark, A. & Thornton, S. 1997. Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences* 20:57-66.

Cliff, D. 1990. Computational neuroethology: A provisional manifesto. Manuscript.

Criticizes connectionism for not being sufficiently rooted in neuroscience, and for not being grounded in the world.

Dawson, M.R.W. & Schopflocher, D.P. 1992. Autonomous processing in parallel distributed processing networks. *Philosophical Psychology* 5:199-219.

Hanson, S. & Burr, D. 1990. What connectionist models learn. *Behavioral and Brain Sciences*.

What's new to connectionism is not learning or representation but the way that learning and representation interact.

Kaplan, S., Weaver, M. & French, R.M. 1990. Active symbols and internal models: Towards a cognitive connectionism. *AI and Society*.

Addresses behaviorist/associationist charges. Connectionism needs recurrent circuits to support active symbols.

Kirsh, D. 1987. Putting a price on cognition. *Southern Journal of Philosophy Supplement* 26:119-35.

On the importance of variable binding, and why it's hard with connectionism.

Lachter, J. & Bever, T. 1988. The relation between linguistic structure and associative theories of language learning. *Cognition* 28:195-247.

Criticism of connectionist language models. They build in too much.

Mills, S. 1989. Connectionism, the classical theory of cognition, and the hundred step constraint. *Acta Analytica* 4:5-38.

Nelson, R. 1989. Philosophical issues in Edelman's neural darwinism. *Journal of Experimental and Theoretical Artificial Intelligence* 1:195-208.

On the relationship between ND, PDP and AI. All are computational.

Oaksford, M., Chater, N. & Stenning, K. 1990. Connectionism, classical cognitive science and experimental psychology. *AI and Society*.

Connectionism is better at explaining empirical findings about mind.

Pinker, S. & Prince, A. 1988. On language and connectionism. *Cognition* 28:73-193.

Extremely thorough criticism of the R&M past-tense-learning model, with arguments on why connectionism can't handle linguistic rules.

4.4 Dynamical Systems

Bechtel, W. 1996. Yet another revolution? Defusing the dynamical system theorists' attack on mental representations. Manuscript.

Clark, A. 1998. Time and mind. *Journal of Philosophy* 95:354-76.

Eliasmith, C. 1996. The third contender: A critical examination of the dynamicist theory of cognition. *Philosophical Psychology* 9:441-63.

Eliasmith, C. 1997. Computation and dynamical models of mind. *Minds and Machines* 7:531-41.

Foss, J.E. 1992. Introduction to the epistemology of the brain: Indeterminacy, micro-specificity, chaos, and openness. *Topoi* 11:45-57.

On the brain as a vector-processing system, and the problems raised by indeterminacy, chaos, and so on. With morals for cognitive science.

Freeman, W. 1997. Nonlinear neurodynamics of intentionality. *Journal of Mind and Behavior* 18:291-304.

Garson, J.W. 1995. Chaos and free will. *Philosophical Psychology* 8:365-74.

Garson, J.W. 1996. Cognition poised at the edge of chaos: A complex alternative to a symbolic mind. *Philosophical Psychology* 9:301-22.

Garson, J.W. 1997. Syntax in a dynamic brain. *Synthese* 110:343-355.

Garson, J.W. 1998. Chaotic emergence and the language of thought. *Philosophical Psychology* 11:303-315.

Giunti, M. 1995. Dynamic models of cognition. In (T. van Gelder & R. Port, eds) *_Mind as Motion_*. MIT Press.

Giunti, M. 1996. *_Computers, Dynamical Systems, and the Mind_*. Oxford University Press.

Globus, G. 1992. Toward a noncomputational cognitive science. *Journal of Cognitive Neuroscience* 4:299-310.

Hooker, C.A. & Christensen, W.D. 1998. Towards a new science of the mind: Wide content and the metaphysics of organizational properties in nonlinear dynamic models. *Mind and Language* 13:98-109.

Horgan, T. & Tienson, J. 1992. Cognitive systems as dynamic systems. *Topoi* 11:27-43.

Horgan, T. & Tienson, J. 1994. A nonclassical framework for cognitive science. *Synthese* 101:305-45.

Keijzer, F.A. & Bem, S. 1996. Behavioral systems interpreted as autonomous agents and as coupled dynamical systems: A criticism. *Philosophical Psychology* 9:323-46.

Slovan, A. 1993. The mind as a control system. In (C. Hookway & D. Peterson, eds) *Philosophy and Cognitive Science*. Cambridge University Press.

van Gelder, T. & Port, R. 1995. *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press.

van Gelder, T. 1995. What might cognition be if not computation? *Journal of Philosophy* 92:345-81.

Argues for a dynamic-systems conception of the mind that is non-computational and non-representational. Uses an analogy with the Watt steam governor to argue for a new kind of dynamic explanation.

van Gelder, T. 1997. Connectionism, dynamics, and the philosophy of mind. In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.

van Gelder, T. 1998. The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* 21:615-28.

4.5 Foundational Questions in AI

4.5a The Nature of AI

Buchanan, B. 1988. AI as an experimental science. In (J. Fetzer, ed) *Aspects of AI*. D. Reidel.

Bundy, A. 1990. What kind of field is AI? In (D. Partridge & Y. Wilks, eds) *The Foundations of Artificial Intelligence: A Sourcebook*. Cambridge University Press.

Dennett, D.C. 1978. AI as philosophy and as psychology. In (M. Ringle, ed) *Philosophical Perspectives on Artificial Intelligence*. Humanities Press. Reprinted in *Brainstorms* (MIT Press, 1978).

AI as detailed armchair psychology and as thought-experimental epistemology. Implications for mind: e.g. a solution to the problem of homuncular regress.

Glymour, C. 1988. AI is philosophy. In (J. Fetzer, ed) *Aspects of AI*. D. Reidel.

Kukla, A. 1989. Is AI an empirical science? *Analysis* 49:56-60.

No, AI is an a priori science that uses empirical methods; its status is similar to that of mathematics.

Kukla, A. 1994. Medium AI and experimental science. *Philosophical Psychology* 7:493-5012.

On the status of "medium AI", the study of intelligence in computational systems (not just humans). Contra to many, this is not an empirical science, but a combination of (experimental) mathematics and engineering.

Nakashima, H. 1999. AI as complex information processing. *Minds and Machines* 9:57-80.

4.5b Levels of Analysis (Marr, etc)

Bechtel, W. 1994. Levels of description and explanation in cognitive science.

Minds and Machines 4:1-25.

Cleeremans, A. & French, R.M. 1996. From chicken squawking to cognition: Levels of description and the computational approach in psychology. *Psychologica Belgica* 36:5-29.

Foster, C. 1990. *_Algorithms, abstraction and implementation_*. Academic Press.

Outlines a theory of the equivalence of algorithms.

Horgan, T. & Tienson, J. 1992. Levels of description in nonclassical cognitive science. *Philosophy* 34, Supplement.

Generalizes Marr's levels to: cognitive state-transitions, mathematical state-transitions, implementation. Discusses these with respect to connectionism, dynamical systems, and computation below the cognitive level.

Houng, Y. 1990. Classicism, connectionism and the concept of level. Dissertation, Indiana University.

On levels of organization vs. levels of analysis.

Marr, D. 1982. *_Vision_*. Freeman.

Defines computational, algorithmic and implementational levels.

McClamrock, R. 1990. Marr's three levels: a re-evaluation. *Minds and Machines* 1:185-196.

On different kinds of level-shifts: organizational and contextual changes. There are more than three levels available.

Newell, A. 1982. The knowledge level. *Artificial Intelligence* 18:81-132.

Newell, A. 1986. The symbol level and the knowledge level. In (Z. Pylyshyn & W. Demopolous, eds) *_Meaning and Cognitive Structure_*. Ablex.

With commentaries by Smith, Dennett.

Peacocke, C. 1986. Explanation in computational psychology: Language, perception and level 1.5. *Mind and Language* 1:101-23.

Psychological explanation is typically somewhere *between* the computational and algorithmic levels.

Sticklen, J. 1989. Problem-solving architectures at the knowledge level. *Journal of Experimental and Theoretical Artificial Intelligence* 1:233-247.

4.5c The Frame Problem

Dennett, D.C. 1984. Cognitive wheels: The frame problem of AI. In (Hookaway, ed) *_Minds, Machines and Evolution_*. Cambridge University Press.

General overview.

Dreyfus, H.L. & Dreyfus, S. 1987. How to stop worrying about the frame problem even though it's computationally insoluble. In (Z. Pylyshyn, ed) *_The Robot's Dilemma_*. Ablex.

FP is an artifact of computational explicitness. Contrast human commonsense know-how, with relevance built in. Comparison to expert/novice distinction.

Fetzer, J.H. 1990. The frame problem: Artificial intelligence meets David Hume. *International Journal of Expert Systems* 3:219-232.

Fodor, J.A. 1987. Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In (Z. Pylyshyn, ed) *_The Robot's Dilemma_*. Ablex.

FP is Hamlet's problem: when to stop thinking. It's equivalent to the general problem of non-demonstrative inference.

Haugeland, J. 1987. An overview of the frame problem. In (Z. Pylyshyn, ed) The Robot's Dilemma. Ablex.

The FP may be a consequence of the explicit/implicit rep distinction. Use "complicit" reps instead, and changes will be carried along intrinsically.

Hayes, P. 1987. What the frame problem is and isn't. In (Z. Pylyshyn, ed) The Robot's Dilemma. Ablex.

FP is a relatively narrow problem, Some, e.g. Fodor, wrongly equate FP with the "Generalized AI Problem".

Janlert, L. 1987. Modeling change: The frame problem. In (Z. Pylyshyn, ed) The Robot's Dilemma. Ablex.

Korb, K. 1998. The frame problem: An AI fairy tale. *Minds and Machines* 8:317-351.

Lormand, E. 1990. Framing the frame problem. *Synthese* 82:353-74.

Criticizes Dennett's, Haugeland's and Fodor's construals of the FP.

Maloney, J.C. 1988. In praise of narrow minds. In (J. Fetzer, ed) Aspects of AI. D. Reidel.

McCarthy, J. & Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. In (Meltzer & Michie, eds) Machine Intelligence 4. Edinburgh University Press.

McDermott, D. 1987. We've been framed: Or, Why AI is innocent of the frame problem. In (Z. Pylyshyn, ed) The Robot's Dilemma. Ablex.

Solve frame problem by using the sleeping-dog strategy -- keeping things fixed unless there's a reason to suppose otherwise.

Pollock, J.L. 1997. Reasoning about change and persistence: A solution to the frame problem. *Nous* 31:143-169.

Pylyshyn, Z.W. (ed) 1987. The Robot's Dilemma. Ablex.

Lots of papers on the frame problem.

4.5d AI Methodology

Birnbaum, L. 1991. Rigor mortis: A response to Nilsson's 'Logic and artificial intelligence'. *Artificial Intelligence* 47:57-78.

Chalmers, D.J., French, R.M. & Hofstadter, D.R. 1992. High-level perception, representation, and analogy: A critique of AI methodology. *Journal of Experimental and Theoretical Artificial Intelligence*.

AI must integrate perception and cognition in the interest of flexible representation. Current models ignore perception and the development of representation, but this cannot be separated from later cognitive processes.

Clark, A. 1986. A biological metaphor. *Mind and Language* 1:45-64.

AI should look at biology.

Clark, A. 1987. The kludge in the machine. *Mind and Language* 2:277-300.

Dascal, M. 1992. Why does language matter to artificial intelligence? *Minds and Machines* 2:145-174.

Dreyfus, H.L. 1981. From micro-worlds to knowledge: AI at an impasse. In (J. Haugeland, ed) Mind Design. MIT Press.

Micro-worlds don't test true understanding, and frames and scripts are doomed to leave out too much. Explicit representation can't capture intelligence.

Dreyfus, H.L. & Dreyfus, S.E. 1988. Making a mind versus modeling the brain: AI at a crossroads. *Daedalus*.

History of AI (boo) and connectionism (qualified hooray). And Husserl/Heidegger/Wittgenstein. Quite nice.

Hadley, R.F. 1991. The many uses of 'belief' in AI. *Minds and Machines* 1:55-74.

Various AI approaches to belief: syntactic, propositional/meaning-based, information, tractability, discoverability, and degree of confidence.

Haugeland, J. 1979. Understanding natural language. *Journal of Philosophy* 76:619-32. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

AI will need holism: interpretational, common-sense, situational, existential.

Kirsh, D. 1991. Foundations of AI: The big issues. *Artificial Intelligence* 47:3-30.

Identifying the dividing lines: pre-eminence of knowledge, embodiment, language-like kinematics, role of learning, uniformity of architecture.

Marr, D. 1977. Artificial intelligence: A personal view. *Artificial Intelligence* 9:37-48.

AI usually comes up with Type 2 (algorithmic) theories, when Type 1 (info processing) theories might be more useful -- at least if they exist.

McDermott, D. 1981. Artificial intelligence meets natural stupidity. In (J. Haugeland, ed) *Mind Design*. MIT Press.

Problems in AI methodology: wishful mnemonics, oversimplifying natural language concepts, and never implementing programs. Entertaining.

McDermott, D. 1987. A critique of pure reason. *Computational Intelligence* 3:151-60.

Criticism of logicism (i.e. reliance on deduction) in AI.

Nilsson, N. 1991. Logic and artificial intelligence. *Artificial Intelligence* 47:31-56.

Partridge, D. & Wilks, Y. (eds) 1990. *The Foundations of Artificial Intelligence: A Sourcebook*. Cambridge University Press.

Lots of papers on various aspects of AI methodology. Quite thorough.

Preston, B. 1993. Heidegger and artificial intelligence. *Philosophy and Phenomenological Research* 53:43-69.

On the non-represented background to everyday activity, and environmental interaction in cognition. Criticizes cognitivism, connectionism, looks at Agre/Chapman/Brooks, ethology, anthropology for support.

Pylyshyn, Z.W. 1979. Complexity and the study of artificial and human intelligence. In (M. Ringle, ed) *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

Ringle, M. (ed) 1979. *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

10 papers on philosophy of AI, psychology and knowledge representation.

Robinson, W.S. 1991. Rationalism, expertise, and the Dreyfuses' critique of AI research. *Southern Journal of Philosophy* 29:271-90.

Defending limited rationalism: i.e. a theory of intelligence below the conceptual level but above the neuronal level.

4.6 Computationalism in Cognitive Science [see also 2.2e]

Antony, L. 1997. Feeling fine about the mind. *Philosophy and Phenomenological Research* 57:381-87.

Bickhard, M. 1996. Troubles with computationalism. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Block, N. 1990. The computer model of mind. In (D. Osherson & E. Smith, eds) *An Invitation to Cognitive Science*, Vol. 3. MIT Press.

Overview of computationalism. Relationship to intentionality, LOT, etc.

Boden, M. 1984. What is computational psychology? *Proceedings of the Aristotelian Society* 58:17-35.

Bringsjord, S. 1994. Computation, among other things, is beneath us. *Minds and Machines* 4:469-88.

Bringsjord, S. & Zenzen, M. 1997. Cognition is not computation: The argument from irreversibility. *Synthese* 113:285-320.

Buller, D.J. 1993. Confirmation and the computational paradigm, or, why do you think they call it artificial intelligence? *Minds and Machines* 3:155-81.

Chalmers, D.J. 1994. A computational foundation for the study of cognition. Manuscript.

Argues for theses of computational sufficiency and computational explanation, resting on the fact that computation provides an abstract specification of causal organization. With replies to many anti-computationalist worries.

Clarke, J. 1972. Turing machines and the mind-body problem. *British Journal for the Philosophy of Science* 23:1-12.

Cummins, R. 1977. Programs in the explanation of behavior. *Philosophy of Science* 44:269-87.

Demopoulos, W. 1987. On some fundamental distinctions of computationalism. *Synthese* 70:79-96.

On analog/digital, representational/nonrepresentational, direct/indirect.

Dietrich, E. 1990. Computationalism. *Social Epistemology*.

What computationalism is, as opposed to computerism & cognitivism. Implies: intentionality isn't special, and we don't make decisions. With commentary.

Dietrich, E. 1989. Semantics and the computational paradigm in computational psychology. *Synthese* 79:119-41.

Argues that computational explanation requires the attribution of semantic content. Addresses Stich's arguments against content, and argues that computers are not formal symbol manipulators.

Double, R. 1987. The computational model of the mind and philosophical functionalism. *Behaviorism* 15:131-39.

Dretske, F. 1985. Machines and the mental. *Proceedings and Addresses of the American Philosophical Association* 59:23-33.

Machines can't even add, let alone think, as the symbols they use aren't meaningful to them. They would need real information based on perceptual embodiment, and conceptual capacities, for meaning to play a real role.

Fetzer, J.H. 1994. Mental algorithms: Are minds computational systems? *Pragmatics and Cognition* 21:1-29.

Fodor, J.A. 1978. Computation and reduction. *Minnesota Studies in the Philosophy of Science* 9. Reprinted in *RePresentations* (MIT Press, 1980).

Fodor, J. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. MIT Press.

Garson, J.W. 1993. Mice in mirrored mazes and the mind. *Philosophical Psychology* 6:123-34.

Computationalism is false, as it can't distinguish the ability to solve a maxe for the ability to solve it's mirror image. An embodied computational taxonomy is needed, rather than software alone.

Harnad, S. 1996. Computation is just interpretable symbol manipulation; Cognition isn't. *Minds and Machines* 4:379-90.

Horst, S. 1996. *Symbols, Computation, and Intentionality: A Critique of the Computational Theory of Mind*. University of California Press.

Horst, S. 1999. Symbols and computation: A critique of the computational theory of mind. *Minds and Machines* 9:347-381

Mellor, D.H. 1984. What is computational psychology? II. *Proceedings of the Aristotelian Society* 58:37-53.

Mellor, D.H. 1989. How much of the mind is a computer. In (P. Slezak, ed) *Computers, Brains and Minds*. Kluwer.

Only belief is computational: rest of mind is not.

Nelson, R. 1987. Machine models for cognitive science. *Philosophy of Science* Argues contra Pylyshyn 1984 that finite state automata are good models for cognitive science: they are semantically interpretable and process symbols.

Pollock, J. 1989. *How to Build a Person: A Prolegomenon*. MIT Press.

Pylyshyn, Z.W. 1980. Computation and cognition: Issues in the foundation of cognitive science. *Behavioral and Brain Sciences* 3:111-32.

Pylyshyn, Z.W. 1984. *Computation and Cognition*. MIT Press.
A thorough account of the symbolic/computational view of cognition.

Pylyshyn, Z.W. 1978. Computational models and empirical constraints. *Behavioral and Brain Sciences* 1:98-128.

Pylyshyn, Z.W. 1989. Computing and cognitive science. In (M. Posner, ed) *Foundations of Cognitive Science*. MIT Press.
An overview of the computational view of mind. On symbols, levels, control structures, levels of correspondence for computational models, and empirical methods for determining degrees of equivalence.

Shapiro, S.C. 1995. Computationalism. *Minds and Machines* 5:467-87.

Sterelny, K. 1989. Computational functional psychology: problems and prospects. In (P. Slezak, ed) *Computers, Brains and Minds*. Kluwer.
Various points on pros and cons of computational psychology.

Tibbetts, P. 1996. Residual dualism in computational theories of mind. *Dialectica* 50:37-52.

4.7 Computation and Physical Systems

Boyle, C.F. 1994. Computation as an intrinsic property. *Minds and Machines* 4:451-67.

Chalmers, D.J. 1994. On implementing a computation. *Minds and Machines* 4:391-402.

Gives an account of what it is for a physical system to implement a

computation: the causal structure of the system must mirror the formal structure of the computation. Answers objections by Searle and others.

Chalmers, D.J. 1996. Does a rock implement every finite-state automaton? *Synthese* 108:309-33.

Argues that Putnam's "proof" that every ordinary open system implements every finite automaton is fallacious. It can be patched up, but an appropriate account of implementation resists these difficulties.

Chrisley, R.L. 1994. Why everything doesn't realize every computation. *Minds and Machines* 4:403-20.

Cleland, C. 1993. Is the Church-Turing thesis true? *Minds and Machines* 3:283-312.

Many physically realized functions can't be computed by Turing machines: e.g. "mundane procedures" and continuous functions. So the C-T thesis is false of these, and maybe even of number-theoretic functions.

Cleland, C.E. 1995. Effective procedures and computable functions. *Minds and Machines* 5:9-23.

Copeland, B.J. 1996. What is computation? *Synthese* 108:335-59.

Endicott, R.P. 1996. Searle, syntax, and observer-relativity. *Canadian Journal of Philosophy* 26:101-22.

Goel, V. 1991. Notationality and the information processing mind. *Minds and Machines* 1:129-166.

Adapts Goodman's notational systems to explicate computational information processing. What is/isn't a physical notational system (e.g. LOT, symbol systems, connectionism) and why. How to reconcile notational/mental content?

Hardcastle, V.G. 1995. Computationalism. *Synthese* 105:303-17.

Pragmatic factors are vital in connecting the theory of computation with empirical theory, and particularly in determining whether a given system counts as performing a given computation.

Horsten, L. 1995. The Church-Turing thesis and effective mundane procedures. *Minds and Machines* 5:1-8.

MacLennan, B. 1994. "Words lie in our way". *Minds and Machines* 4:421-37.

Miscevic, N. 1996. Computationalism and the Kripke-Wittgenstein paradox. *Proceedings of the Aristotelian Society* 96:215-29.

Scheutz, M. 1999. When physical systems realize functions. *Minds and Machines* 9:161-196.

Searle, J.R. 1990. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association* 64:21-37.

Syntax isn't intrinsic to physics, so computational ascriptions are assigned by observer. Syntax has no causal powers. Brain doesn't process information.

Shagrir, O. 1997. Two dogmas of computationalism. *Minds and Machines* 7:321-44.

Stabler, E. 1987. Kripke on functionalism and automata. *Synthese* 70:1-22.

Disputes Kripke's argument that there is no objective way of determining when a system computes a given function, due to infinite domains and unreliability. Stipulating normal background conditions is sufficient.

Suber, P. 1988. What is software? *Journal of Speculative Philosophy* 2:89-119.

4.8 Philosophy of AI, Misc

Bergadano, F. 1993. Machine learning and the foundations of inductive inference. *Minds and Machines* 3:31-51.

Button, G., Coulter, J., Lee, J.R.E. & Sharrock, W. 1995. *_Computers, Minds, and Conduct_*. Polity Press.

Fetzer, J.H. 1990. *_Artificial Intelligence: Its Scope and Limits_*. Kluwer.

Gips, J. 1994. Toward the ethical robot. In (K.M. Ford, C. Glymour, & P. Hayes, eds) *_Android Epistemology_*. MIT Press.

Haugeland, J. (ed) 1981. *_Mind Design_*. MIT Press.
12 papers on the foundations of AI and cognitive science.

Hayes, P.J., Ford, K.M., & Adams-Webber, J.R. 1994. Human reasoning about artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence* 4:247-63. Reprinted in (E. Dietrich, ed) *_Thinking Computers and Virtual Persons_*. Academic Press.

Krellenstein, M. 1987. A reply to 'Parallel computation and the mind-body problem'. *Cognitive Science* 11:155-7.
Thagard 1986 is wrong: speed and the like make no fundamental difference.
With Thagard's reply: it makes a difference in practice, if not in principle.

Moody, T.C. 1993. *_Philosophy and Artificial Intelligence_*. Prentice-Hall.

Preston, B. 1991. AI, anthropocentrism, and the evolution of "intelligence.". *Minds and Machines* 1:259-277.

Robinson, W.S. 1992. *_Computers, Minds, and Robots_*. Temple University Press.

Russell, S. 1991. Inductive learning by machines. *Philosophical Studies* 64:37-64.
A nice paper on the relationship between techniques of theory formation from machine learning and philosophical problems of induction and knowledge.

Rychlak, J.F. 1991. *_Artificial Intelligence and Human Reason: A Teleological Critique_*. Columbia University Press.

Sloman, A. 1978. *_The Computer Revolution in Philosophy_*. Harvester.
All about how the computer should change the way we think about the mind.

Thagard, P. 1986. Parallel computation and the mind-body problem. *Cognitive Science* 10:301-18.
Parallelism does make a difference. Some somewhat anti-functionalist points.

Thagard, P. 1990. Philosophy and machine learning. *Canadian Journal of Philosophy* 20:261-76.

Thagard, P. 1991. Philosophical and computational models of explanation. *Philosophical Studies* 64:87-104.
A comparison of philosophical and AI approaches to explanation: deductive, statistical, schematic, analogical, causal, and linguistic.

Winograd, T. & Flores, F. 1987. *_Understanding Computers and Cognition_*. Addison-Wesley.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 5: Philosophy of Psychology [561]

=====

Contents

- 5.1 Issues in Cognitive Science [230]
 - 5.1a Nativism (Chomsky, etc) [37]
 - 5.1b Modularity (Fodor, etc) [30]
 - 5.1c Mental Imagery [45]
 - 5.1d Rationality [31]
 - 5.1e Embodiment [31]
 - 5.1f Animal Cognition [56]
- 5.2 Aspects of Mind [224]
 - 5.2a Pain and Pleasure [40]
 - 5.2b Emotions [46]
 - 5.2c Dreams [21]
 - 5.2d Memory [32]
 - 5.2e Color [85]
- 5.3 Philosophy of Psychology and Cognitive Science [80]
 - 5.3a Psychological Laws [13]
 - 5.3b Psychology and Neuroscience [33]
 - 5.3c Explanation in Cognitive Science [17]
 - 5.3d Philosophy of Cognitive Science, Misc [17]
- 5.4 Philosophy of Mind (General) [27]

5.1 Issues in Cognitive Science

5.1a Nativism (Chomsky, etc)

Ariew, A. 1996. Innateness and canalization. *Philosophy of Science Supplement* 63:19-27.

Atherton, M. & Schwarz, R. 1974. Linguistic innateness and its evidence. *Journal of Philosophy* 71:6.

Chomsky, N. 1967. Recent contributions to the theory of innate ideas. *Synthese* 17:2-11.

Chomsky, N. 1969. Linguistics and philosophy. In (S. Hook, ed) *_Language and Philosophy_*. New York University Press.

Reply to Putnam 1967: Putnam underestimates complexity of grammar, etc.

Chomsky, N. 1975. On cognitive capacity. In *_Reflections on Language_*. Pantheon Books.

Chomsky, N. 1980. Discussion of Putnam's comments. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

Chomsky, N. & Fodor, J.A. 1980. The inductivist fallacy. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

Churchland, P.S. 1978. Fodor on language learning. *Synthese* 38:149-59.

Cowie, F. 1997. The logical problem of language acquisition. *Synthese* 111:17-51.

Cowie, F. 1998. *_What's Within?_* Oxford University Press.

Cowie, F. 1998. Mad dog nativism. *British Journal for the Philosophy*

of Science 49:227-252.

Cummins, D.D. 1996. Evidence for the innateness of deontic reasoning. *Mind and Language* 11:160-90.

De Rosa, R. 2000. On Fodor's claim that classical empiricists and rationalists agree on the innateness of ideas. *Protosociology* 14:240-269.

Harman, G. 1969. Linguistic competence and empiricism. In (S. Hook, ed) *_Language and Philosophy_*. New York University Press.

Fodor, J.A., Bever, T. & Garrett, M. 1974. The specificity of language skills. In *_The Psychology of Language_*. McGraw-Hill.

Fodor, J.A. 1980. Reply to Putnam. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

Fodor, J.A. 1981. The present status of the innateness controversy. In *_Representations_*. MIT Press.
Concepts are undefinable, so primitive, so innate (plus gloss).

Fodor, J.A. 1980. On the impossibility of acquiring 'more powerful' structures. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

Katz, J. 1966. Innate ideas. In *_The Philosophy of Language_*. Harper & Row.
Overview; poverty of stimulus, unobservable features => rationalism.

Kaye, L.J. 1993. Are most of our concepts innate? *Synthese* 2:187-217.

Mehler, J. & Fox, R. (eds) 1985. *_Neonate Cognition: Beyond the Blooming Buzzing Confusion_*. Lawrence Erlbaum.

Piattelli-Palmarini, M. (ed) 1980. *_Language and Learning: The Debate Between Jean Piaget and Noam Chomsky_*. Harvard University Press.
An excellent collection of papers & responses by Piaget, Chomsky and others.

Piattelli-Palmarini, M. 1986. The rise of selective theories: A case study and some lessons from immunology. In (W. Demopoulos, ed) *_Language Learning and Concept Acquisition_*. Ablex.

Piattelli-Palmarini, M. 1989. Evolution, selection, and cognition: From learning to parameter setting in biology and in the study of language. *Cognition* 31:1-44.
Why learning is selective and not instructive. Biological analogies, linguistic evidence. Dispense with "learning" as a scientific term.

Pitt, D. 2000. Nativism and the theory of content. *Protosociology* 14:222-239.

Putnam, H. 1967. The 'Innateness Hypothesis' and explanatory models in linguistics. *Synthese* 17:12-22. Reprinted in *_Mind, Language, and Reality_* (Cambridge University Press, 1975).
Contra nativism: disputes (1) surprising universals (2) explanation of universals (3) ease of learning (4) relevance of IQ-independence.

Putnam, H. 1980. What is innate and why. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

Putnam, H. 1980. Comments on Chomsky's and Fodor's replies. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

- Ramsey, W. & Stich, S.P. 1990. Connectionism and three levels of nativism. *Synthese* 82:177-205.
Identifies minimal nativism vs anti-empiricism vs rationalism. Considers the relevance of connectionist networks. Some nativist arguments may survive.
- Samet, J. 1986. Troubles with Fodor's nativism. *Midwest Studies in Philosophy* 10:575-594.
Concepts can be acquired without being learned by symbol-manipulation.
- Samet, J. & Flanagan, O.J. 1989. Innate representations. In (S. Silvers, ed) *_Rerepresentation_*. Kluwer.
- Sampson, G. 1978. Linguistic universals as evidence for empiricism. *Journal of Linguistics*.
Explain universals via Popper/Simon empirical model.
- Samuels, R. 1998. What brains won't tell us about the mind: A critique of the neurobiological argument against representational nativism. *Mind and Language* 13:548-570.
- Schwartz, R. 1995. Is mathematical competence innate? *Philosophy of Science* 62:227-40.
- Sterelny, K. 1989. Fodor's nativism. *Philosophical Studies* 55:119-41.
- Stich, S.P. (ed) 1975. *_Innate Ideas_*. University of California Press.
- Stich, S.P. 1979. Between Chomskian rationalism and Popperian empiricism. *British Journal for the Philosophy of Science* 30:329-47.
Can take middle ground. Anti-empiricism doesn't imply rationalism.

5.1b Modularity (Fodor, etc)

- Appelbaum, I. 1998. Fodor, modularity, and speech perception. *Philosophical Psychology* 11:317-330.
- Arbib, M. 1989. Modularity, schemas and neurons: A critique of Fodor. In (P. Slezak, ed) *_Computers, Brains and Minds_*. Kluwer.
Against Fodor: modules are smaller, interact strongly, not domain-specific.
- Bennett, L.J. 1990. Modularity of mind revisited. *British Journal for the Philosophy of Science* 41:429-36.
Remarks on Shanon and Fodor.
- Browne, D. 1996. Cognitive versatility. *Minds and Machines* 6:507-23.
- Bruner, J. 1957. On perceptual readiness. *Psychological Review* 65:14-21.
Overview of the original studies on top-down effects in perception.
- Cam, P. 1988. Modularity, rationality, and higher cognition. *Philosophical Studies* 53:279-94.
- Cam, P. 1990. Insularity and the persistence of perceptual illusion. *Analysis* 50:231-5.
- Chien, A.J. 1996. Why the mind may not be modular. *Minds and Machines* 6:1-32.
- Churchland, P.M. 1979. *_Scientific Realism and the Plasticity of Mind_*. Cambridge University Press.
Our perception is deeply theory-laden, and potentially very plastic.
- Churchland, P.M. 1988. Perceptual plasticity and theoretical neutrality: A

reply to Jerry Fodor. *Philosophy of Science* 55:167-87. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).

Contra Fodor 1984: observation is theory-laden (built-in or not); supported by neurophysiological evidence; perceptual systems have long-term plasticity.

Currie, G. & Sterelny, K. 2000. How to think about the modularity of mind-reading. *Philosophical Quarterly* 50:145-160.

DesAutels, P. 1995. Two types of theories: The impact of Churchland's perceptual plasticity. *Philosophical Psychology* 8:25-33.

Fodor, J.A. 1983. *_The Modularity of Mind_*. MIT Press.
Perception happens in informationally encapsulated, domain-specific modules.
Central systems aren't encapsulated, and so may be impossible to understand.

Fodor, J.A. 1985. *Precis of _The modularity of mind_*. *Behavioral and Brain Sciences* 8:1-42. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

Summary of MOM (with commentary and reply in the BBS printing).

Fodor, J.A. 1986. The modularity of mind. In (Z. Pylyshyn, ed) *_Meaning and Cognitive Structure_*. Ablex.

Informal discussion of modularity. With commentaries by Fahlman, Caplan.

Fodor, J.A. 1984. Observation reconsidered. *Philosophy of Science* 51:23-43. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

Argues for an observation/theory distinction, and against belief affecting perception.

Fodor, J.A. 1988. A reply to Churchland's 'Perceptual plasticity and theoretical neutrality'. *Philosophy of Science* 55:188-98. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

Churchland is up the creek without a paddle.

Fodor, J.A. 1989. Why should the mind be modular? In (A. George, ed) *_Reflections on Chomsky_*. Blackwell. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

Garfield, J. (ed) 1987. *_Modularity in Knowledge Representation and Natural-Language Understanding_*. MIT Press.

A collection of papers on modularity in language and vision.

Gray, R. 2001. Cognitive modules, synaesthesia and the constitution of psychological natural kinds. *Philosophical Psychology* 14:65-82.

Meyering, T.C. 1994. Fodor's modularity: A new name for an old dilemma. *Philosophical Psychology* 7:39-62.

Olsson, E. 1997. Coherence and the modularity of mind. *Australasian Journal of Philosophy* 75:404-11.

Pylyshyn, Z. 1999. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences* 22:341-365.

Rollins, M. 1994. Deep plasticity: The encoding approach to perceptual change. *Philosophy of Science* 61:39-54.

Ross, J. 1990. Against postulating central systems in the mind. *Philosophy of Science* 57:297-312.

Fodor's arguments for unencapsulated central systems are no good; AI is possible after all.

Shanon, B. 1988. Remarks on the modularity of mind. *British Journal for the Philosophy of Science* 39:331-52.

Criticism of Fodor. Modularity is dynamic, and can be central.

Fodor, J.A. & Pylyshyn, Z.W. 1981. How direct is visual perception?: Some reflections on Gibson's 'ecological approach'. *Cognition* 9:139-96.
'Direct perception' can't correspond to anything. Perception is inferential.

Turvey, M.T., Shaw, R.E., Reed, E.S., & Mace, W.M. 1981. Ecological laws of perceiving and acting: In Reply to Fodor and Pylyshyn. *Cognition* 9:237-304.

Ullman, S. 1980. Against direct perception. *Behavioral and Brain Sciences* 3:333-81.

Vaina, L.M. 1990. What and where in the human visual system: Two hierarchies of visual modules. *Synthese* 83:49-91.

5.1c Mental Imagery [see also 6.2j]

Anderson, J.R. 1978. Arguments concerning representations for mental imagery. *Psychological Review* 85.

Audi, R. 1978. The ontological status of mental images. *Inquiry* 21:348-61.

Blachowicz, J. 1997. Analog representation beyond mental imagery. *Journal of Philosophy* 94:55-84.

Block, N. (ed) 1981. *_Imagery_*. MIT Press.

Block, N. 1983. Mental pictures and cognitive science. *Philosophical Review* 93:499-542. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Block, N. 1983. The photographic fallacy in the debate about mental imagery. *Nous* 17:651-62.

Brown, R. & Herrnstein, R. 1981. Icons and images. In (N. Block, ed) *_Imagery_*. MIT Press.

Cam, P. 1987. Propositions about images. *Philosophy and Phenomenological Research* 48:335-8.

Candlish, S. 1975. Mental images and pictorial properties. *Mind* 84:260-2.

Chambers, D. & Reisberg, D. 1985. Can mental images be ambiguous?' *Journal of Experimental Psychology: Human Perception and Performance* 11:317-28.

Chambers, D. & Reisberg, D. 1992. What an image depicts depends on what an image means. *Cognitive Psychology* 24:145-74.

Dennett, D.C. 1978. Two approaches to mental images. In *_Brainstorms_*. MIT Press.

Dennett, D.C. 1968. The nature of images and the introspective trap. In *_Content and Consciousness_*. Routledge and Kegan Paul. Reprinted in (N. Block, ed) *_Readings in the Philosophy of Psychology_* (MIT Press, 1980).

Farah, M.J. 1988. Is visual imagery really visual: Some overlooked evidence from neuropsychology. *Psychological Review* 95:307-17.

Finke, R.A. 1989. *_Principles of Mental Imagery_*. MIT Press.

Fodor, J.A. 1975. Imagistic representation. In *_The Language of Thought_*.

Harvard University Press.

Glasgow, J.I. 1993. The imagery debate revisited: A computational perspective. *Computational Intelligence* 9:310-33.

Hannay, A. 1971. *Mental Images: A Defense*. Allen & Unwin.

Hannay, A. 1973. To see a mental image. *Mind* 82:161-262.

Kind, A. 2001. Putting the image back in imagination. *Philosophy and Phenomenological Research* 62:85-110.

Kosslyn, S.M. & Pomerantz, J. 1977. Imagery, propositions and the form of internal representations. *Cognitive Psychology* 9:52-76. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Kosslyn, S.M. 1981. The medium and the message in mental imagery: A theory. In (N. Block, ed) *Imagery*. MIT Press.

Kosslyn, S.M., Pinker, S., Schwartz, S. & Smith, G. 1979. On the demystification of mental imagery. *Behavioral and Brain Sciences* 2:535-81.

Kosslyn, S.M. 1980. *Image and Mind*. Harvard University Press.

Kosslyn, S.M. 1994. *Image and Brain: The Resolution of the Imagery Debate*. MIT Press.

Maloney, J.C. 1984. Mental images and cognitive theory. *American Philosophical Quarterly* 21:237-47.

Morris, P.E. & Hampson, P.J. 1983. *Imagery and Consciousness*. Academic Press.

Mortensen, C. 1989. Mental images: Should cognitive science learn from neurophysiology? In (P. Slezak, ed) *Computers, Brains and Minds*. Kluwer.

Pylyshyn, Z.W. 1973. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psych Bull* 80:1-24.

Pylyshyn, Z.W. 1978. Imagery and artificial intelligence. In (W. Savage, ed) *Perception and Cognition*. University of Minnesota Press. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Pylyshyn, Z.W. 1981. The imagery debate: Analog media vs. tacit knowledge. *Psychological Review* 88:16-45. Reprinted in Block 1981.

Reisberg, D. & Chambers, D. 1991. Neither pictures nor propositions: What can we learn from a mental image? *Canadian Journal of Psychology* 45:336-52.

Rey, G. 1981. What are mental images? In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol. 2. Harvard University Press.

Richardson, A. 1969. *Mental Imagery*. Routledge.

Rollins, M. 1989. *Mental Imagery: On the Limits of Cognitive Science*. Yale University Press.

Russow, L. 1985. Dennett, mental images and images in context. *Philosophy and Phenomenological Research* 45:581-94.

Schwartz, R. 1980. Imagery: There is more to it than meets the eye. *Philosophy of Science Association* 1980.

Shepard, R. & Cooper, L. 1982. *Mental Images and their Transformations*.

MIT Press.

Shier, D. 1997. How can pictures be propositions? *Ratio* 10:65-75.

Sterelny, K. 1986. The imagery debate. *Philosophy of Science* 53:560-83.
Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Thomas, N.J.T. 1997. Are theories of imagination theories of imagery?
Manuscript.

Tye, M. 1984. The debate about mental imagery. *Journal of Philosophy*
81:678-91.

Tye, M. 1988. The picture theory of images. *Philosophical Review*.

Tye, M. 1991. *The Imagery Debate*. MIT Press,

Wright, E. 1983. Inspecting images. *Philosophy* 58:57-72.

5.1d Rationality

Biro, J. & Ludwig, K. 1994. Are there more than minimal a priori limits on
irrationality? *Australasian Journal of Philosophy* 72:89-102.

Cherniak, C. 1986. *Minimal Rationality*. MIT Press.

Cherniak, C. 1981. Minimal rationality. *Mind* 90:161-83.

Cherniak, C. 1983. Rationality and the structure of memory. *Synthese*
57:163-86.

Cohen, L.J. 1979. On the psychology of prediction: Whose is the fallacy?
Cognition 7:385-407.

Cohen, L.J. 1980. Whose is the fallacy? A rejoinder to Daniel Kahneman and
Amos Tversky. *Cognition* 8:89-92.

Cohen, L.J. 1981. Can human irrationality be experimentally demonstrated?
Behavioral and Brain Sciences.

Cohen, L.J. 1986. *The Dialogue of Reason*. Cambridge University Press.

Cook, K.S. & Levi, M. 1990. *The Limits of Rationality*. University of
Chicago Press.

Davidson, D. 1985. Incoherence and irrationality. *Dialectica* 39:345-54.

Davidson, D. 1995. Could there be a science of rationality? *International
Journal of Philosophical Studies* 3:1-16.

Feldman, R. 1988. Rationality, reliability, and natural selection. *Philosophy
of Science* 55:218-27.

Fetzer, J.H. 1990. Evolution, rationality and testability. *Synthese*
82:423-39.

Gardner, S. 1996. *Irrationality and the Philosophy of Psychoanalysis*.
Cambridge University Press.

Harman, G. 1986. *Change in View*. MIT Press.

Holt, L. 1999. Rationality is still hard work: Some further notes on the
disruptive effects of deliberation. *Philosophical Psychology* 12:215-219.

- Kahneman, D., Slovic, P. & Tversky, A. (eds) 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.
- Kahneman, D. & Tversky, A. 1979. On the interpretation of intuitive probability: A reply to Jonathan Cohen. *Cognition* 7:409-11.
- Manktelow, K. & Over, D. 1987. Reasoning and rationality. *Mind and Language* 2:199-219.
- Mele, A.R. 1987. *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford University Press.
- Nisbett, R. & Ross, L. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall.
- Nozick, R. 1993. *The Nature of Rationality*. Princeton University Press.
- Reiner, R. 1995. Arguments against the possibility of perfect rationality. *Minds and Machines* 5:373-89.
- Rust, J. 1990. Delusions, irrationality, and cognitive science. *Philosophical Psychology*.
- Scholl, B.J. 1997. Reasoning, rationality, and architectural resolution. *Philosophical Psychology* 10:451-470.
- Scott-Kakures, D. 1996. Self-deception and internal irrationality. *Philosophy and Phenomenological Research* 56:31-56.
- Sober, E. 1981. The evolution of rationality. *Synthese* 46:95-120.
- Stein, E. 1994. Rationality and reflective equilibrium. *Synthese* 99:137-72.
- Stein, E. 1996. *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford University Press.
- Stich, S.P. 1985. Could man be an irrational animal? *Synthese* 64:115-35.
- Wason, P. 1966. Reasoning. In (Foss, ed) *New Horizons in Psychology*. Penguin.

5.1e Embodiment [see also 2.2] -----

- Agre, P. 1995. Computation and embodied agency. *Informatica* 19:527-35.
- Ballard, D. 1991. Animate vision. *Artificial Intelligence* 48:57-86.
- Beer, R. 1995. A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* 72:173-215.
- Bermudez, J.L., Marcel, A., & Eilan, N. (eds) 1995. *The Body and the Self*. MIT Press.
- Buckley, J. & Hall, L. 1999. Self-knowledge and embodiment. *Southwest Philosophy Review* 15.
- Chrisley, R.L. 1994. Taking embodiment seriously: Nonconceptual content and robotics. In (K.M. Ford, C. Glymour, & P. Hayes, eds) *Android Epistemology*. MIT Press.
- Clark, A. 1987. Being there: Why implementation matters to cognitive science. *AI Review* 1:231-44.
- On the importance of embodiment of systems in cognition.

Clark, A. 1995. Moving minds: Situating content in the service of real-time success. *Philosophical Perspectives* 9:89-104.

Clark, A. 1997. *_Being There: Putting Brain, Body, and World Together Again_*. MIT Press.

Clark, A. & Chalmers, D.J. 1998. The extended mind. *Analysis* 58:7-19.
Advocates a sort of "active externalism", based on the role of the environment in actively driving cognition. Beliefs can extend into an agent's immediate environment (e.g. a notebook) in this way.

Clark, A. 2001. Reasons, robots and the extended mind. *Mind and Language* 16:121-45.

Cussins, A. 1992. Content, embodiment, and objectivity: The theory of cognitive trails. *Mind* 101:651-88.

Gibson, J.J. 1979. *_The Ecological Approach to Visual Perception_*. Houghton Mifflin.

Godfrey-Smith, P. 1996. *_Complexity and the Function of Mind in Nature_*. Cambridge University Press.

Haugeland, J. 1993. Mind embodied and embedded. In (Y. Houn & J. Ho, eds) *_Mind and Cognition: 1993 International Symposium_*. Academia Sinica.
Argues that the mind is not just embedded but intimately intermingled with the world. With some systems-theoretic arguments arguing against a determinate interface. Mind is not an inner realm.

Hendriks-Jansen, H. 1996. *_Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought_*. MIT Press.

Hutchins, E. 1995. *_Cognition in the Wild_*. MIT Press.

Johnson, M.L. 1987. *_The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason_*. University of Chicago Press.

Johnson, M.L. 1995. Incarnate mind. *Minds and Machines* 5:533-45.

Kirsh, D. & Maglio, P. 1995. On distinguish epistemic from pragmatic action. *Cognitive Science* 18:513-49.

Loren, L.A. & Dietrich, E. 1997. Merleau-Ponty, embodied cognition, and the problem of intentionality. *Cybernetics and Systems* 28:345-58.

Losonsky, M. 1995. Emdedded systems vs. individualism. *Minds and Machines* 5:357-71.

McClamrock, R. 1995. *_Existential Cognition: Computational Minds in the World_*. University of Chicago Press.

O'Regan, K. 1992. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology* 46:461-88.

Rosenschein, S.J. & Kaelbling, L.P. 1995. A situated view of representation and control. *Artificial Intelligence* 73:149-73.

van Gelder, T. 1993. The distinction between mind and cognition. In (Y. Houn & J. Ho, eds) *_Mind and Cognition: 1993 International Symposium_*. Academia Sinica.
Argues against the contemporary "Cartesian" view of mind as an ontologically homogeneous inner representational realm that causes behavior, arguing for a holistic embodied view instead. Mind is therefore safe from elimination.

Varela, F., Thompson, E. & Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Vera, A.H. & Simon, H.A. 1993. Situated action: A symbolic interpretation. *Cognitive Science* 17:7-48.

Wells, A. 1996. Situated action, symbol systems and universal computation. *Minds & Machines* 6:33-46.

Wilkerson, W.S. 1999. From bodily motions to bodily intentions: the perception of bodily activity. *Philosophical Psychology* 12:61-77.

Zhang, J. & Norman, D. 1994. Representations in distributed cognitive tasks. *Cognitive Science* 18:87-122.

5.1f Animal Cognition [see also 6.4c, 6.4d]

Allen, C. 1997. Animal cognition and animal minds. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Allen, C. 1999. Animal concepts revisited: the use of self-monitoring as an empirical approach. *Erkenntnis* 51:537-544.

Allen, C. & Bekoff, M. 1992. On aims and methods of cognitive ethology. *Philosophy of Science Association* 1992, 2:110-24.

Allen, C. & Bekoff, M. 1995. Cognitive ethology and the intentionality of animal behavior. *Mind and Language* 10:313-328.

Allen, C. & Bekoff, M. 1997. *Species of Mind: The Philosophy and Biology of Cognitive Ethology*. MIT Press.

Allen, C. & Hauser, M. 1991. Concept attribution in nonhuman animals: Theoretical and methodological problems in ascribing complex mental processes. *Philosophy of Science* 58:221-40. Reprinted in Allen & Jamison 1996.

Bateson, P.P.G. & Klopfer, P.H. 1991. *Perspectives in Ethology, Volume 9: Human Understanding and Animal Awareness*. Plenum Press.

Beer, C.G. 1992. Conceptual issues in cognitive ethology. *Advances in the Study of Behavior* 21:69-109.

Bekoff, M. & Jamieson, D. (eds) 1996. *Readings in Animal Cognition*. MIT Press.

Bekoff, M. 1999. Social cognition: Exchanging and sharing information on the run. *Erkenntnis* 51:617-632.

Cheney, D.L. & Seyfarth, R.M. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. University of Chicago Press.

Clark, S.R.L. 1987. The description and evaluation of animal emotion. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Cockburn, D. 1994. Human beings and giant squids (on ascribing human sensations and emotions to non-human creatures). *Philosophy* 69:135-50.

Crisp, R. 1996. Evolution and psychological unity. In (M. Bekoff & D. Jamieson, eds) *Readings in Animal Cognition*. MIT Press.

Davidson, D. 1982. Rational animals. *Dialectica* 36:317-28.

Dawkins, M.S. 1987. Minding and mattering. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Dawkins, M.S. 1990. From an animal's point of view: Motivation, fitness, and animal welfare. *Behavioral and Brain Sciences*.

Dennett, D.C. 1983. Intentional systems in cognitive ethology: The 'Panglossian paradigm' defended. *Behavioral and Brain Sciences* 6:343-90. Reprinted in *The Intentional Stance* (MIT Press, 1987).

Dennett, D.C. 1989. Cognitive ethology: Hunting for bargains or a wild goose chase? In (Montefiore, ed) *Goals, No-Goals and Own Goals*. Unwin Hyman.

Dennett, D.C. 1995. Do animals have beliefs? In (H. Roitblat & J. Meyer, eds) *Comparative Approaches to Cognitive Science*. MIT Press.

Dennett, D.C. 1996. *Kinds of Minds*. Basic Books.

Dreckmann, F. 1999. Animal beliefs and their contents. *Erkenntnis* 51:597-615.

Dupre, J. 1996. The mental lives of nonhuman animals. In (M. Bekoff & D. Jamieson, eds) *Readings in Animal Cognition*. MIT Press.

Fellows, R. 2000. Animal belief. *Philosophy* 75:587-599.

Gaita, R. 1992. Animal thoughts. *Philosophical Investigations* 15:227-44.

Gauker, C. 1990. How to learn language like a chimpanzee. *Philosophical Psychology* 4:139-46.

Glock, H. 2000. Animals, thoughts and concepts. *Synthese* 123:35-104.

Gould, J.L. & Gould, C.G. 1982. The insect mind: Physics or metaphysics? In (D. Griffin, ed) *Animal Mind -- Human Mind*. Springer-Verlag.

Gould, J.L. & Gould, C.G. 1994. *The Animal Mind*. Scientific American Library.

Griffin, D.R. (ed) 1982. *Animal Mind -- Human Mind*. Springer-Verlag.

Griffin, D.R. 1984. *Animal Thinking*. Harvard University Press.

Griffin, D.R. 1992. *Animal Minds*. University of Chicago Press.

Harrison, P. 1991. Do animals feel pain? *Philosophy* 66:25-40.

Heil, J. 1982. Speechless brutes. *Philosophy and Phenomenological Research* 42:400-406.

Hendrichs, H. 1999. Different roots of human intentionality in mammalian mentality. *Erkenntnis* 51:649-668.

Malcolm, N. 1973. Thoughtless brutes. *Proceedings and Addresses of the American Philosophical Association* 46:5-20.

Nelson, J. 1983. Do animals propositionally know? Do they propositionally believe? *American Philosophical Quarterly* 20:149-60.

Premack, D. & Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 4:515-629.

Premack, D. 1986. *Gavagai! or the Future History of the Animal Language Controversy*. MIT Press.

Proust, J. 1999. Mind, space and objectivity in non-human animals. *Erkenntnis* 51:545-562.

Radner, D. 1993. Directed action and animal communication. *Ration* 6:135-54.

Radner, D. 1999. Mind and function in animal communication. *Erkenntnis* 51:633-648.

Ristau, C.A. (ed) 1991. *Cognitive Ethology: The Minds of Other Animals*. Lawrence Erlbaum.

Roberts, R.C. 1996. Propositions and animal emotion. *Philosophy* 71:147-56.

Routley, R. 1982. Alleged problems in attributing beliefs, and intentionality, to animals. *Inquiry* 24:385-417.

Savage-Rumbaugh, E.S., Rumbaugh, D.M., & Boysen, S. 1980. Do apes use language? *American Scientist* 68:49-61.

Savage-Rumbaugh, S. & Brakke, K.E. 1996. Animal language: Methodological and interpretative issues. In (C. Allen & D. Jamison, eds) *Readings in Animal Cognition*. MIT Press.

Sebeok, T.A. & Umiker-Sebeok, J. 1980. *Speaking of Apes: A Critical Anthology of Two-Way Communication with Man*. Plenum Press.

Smit, H. 1995. Are animal displays bodily movements or manifestations of the mind? *Behavior and Philosophy* 23:13-19.

Sorabji, R. 1992. Animal minds. *Southern Journal of Philosophy* 31:1-18.

Stephan, A. 1999. Are animals capable of concepts? *Erkenntnis* 51:583-596.

Sterelny, K. 1990. Animals and individualism. In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.

Sterelny, K. 1995. Basic minds. *Philosophical Perspectives* 9:251-70.

Stich, S.P. 1978. Do animals have beliefs? *Australasian Journal of Philosophy* 57:15-28.

Wilder, H. 1996. Interpretative cognitive ethology. In (C. Allen & D. Jamison, eds) *Readings in Animal Cognition*. MIT Press.

Wilson, M.D. 1995. Animal ideas. *Proceedings and Addresses of the American Philosophical Association* 69:7-25.

5.2 Aspects of Mind -----

5.2a Pain and Pleasure -----

Aydede, M. 2000. An analysis of pleasure vis-a-vis pain. *Philosophy and Phenomenological Research* 61:537-570.

Beardman, S. 2000. The choice between current and retrospective evaluations of pain. 13:97-110.

Blum, A. 1991. A note on pleasure. *Journal of Value Inquiry* 25:367-70.

Chapman, C.R. 2000. Pain and folk theory. *Brain and Mind* 1:209-222.

- Conee, E. 1984. A defense of pain. *Philosophical Studies* 46:239-48.
- Cowan, J. 1968. *Pleasure and Pain: A Study in Philosophical Psychology*. Macmillan.
- Dartnall, T. 2001. The pain problem. *Philosophical Psychology* 14:95-102.
- Dennett, D.C. 1978. Why you can't make a computer that feels pain. *Synthese* 38. Reprinted in *Brainstorms* (MIT Press, 1978).
- Douglas, G. 1998. Why pains are not mental objects. *Philosophical Studies* 91:127-148.
- Edwards, R. 1975. Do pleasures and pains differ qualitatively? *Journal of Value Inquiry* 9:270-81.
- Garfield, J.L. 2001. Pain deproblematized. *Philosophical Psychology* 14:103-7.
- Gillett, G. 1991. The neurophilosophy of pain. *Philosophy* 66:191-206.
- Goldstein, I. 1980. Why people prefer pleasure to pain. *Philosophy* 55.
- Goldstein, I. 1989. Pleasure and pain: unconditional intrinsic values. *Philosophy and Phenomenological Research*.
- Goldstein, I. 1999. Intersubjective properties by which we specify pain, pleasure, and other kinds of mental states. *Philosophy*.
- Graham, G. & Stephens, G. 1987. Minding your P's and Q's: Pain and sensible qualities. *Nous* 21:395-405.
- Grahek, N. 1991. Objective and subjective aspects of pain. *Philosophical Psychology* 4:249-66.
- Grahek, N. 1995. The sensory dimension of pain. *Philosophical Studies* 79:167-84.
- Gustafson, D. 1995. Belief in pain. *Consciousness and Cognition* 4:323-45.
- Gustafson, D. 2000. On the supposed utility of a folk theory of pain. *Brain and Mind* 1:223-228.
- Hall, R.J. 1989. Are pains necessarily unpleasant? *Philosophy and Phenomenological Research* 49:643-59.
- Hardcastle, V.G. 1997. When a pain is not. *Journal of Philosophy* 94:381-409.
- Hardcastle, V.G. 2000. *The Myth of Pain*. MIT Press.
- Kaufman, R. 1985. Is the concept of pain incoherent? *Southern Journal of Philosophy* 23:279-84.
- Langsam, H. 1995. Why pains are mental objects. *Journal of Philosophy* 6:303-13.
- Momeyer, R. 1975. Is pleasure a sensation? *Philosophy and Phenomenological Research* 36:113-21.
- Morris, K.J. 1996. Pain, injury, and first/third-person asymmetry. *Philosophy and Phenomenological Research* 56:125-56.
- Nelkin, N. 1986. Pains and pain sensations. *Journal of Philosophy* 83:129-48.
- Nelkin, N. 1994. Reconsidering pain. *Philosophical Psychology* 7:325-43.

- Newton, N. 1989. On viewing pain as a secondary quality. *Nous* 23:569-98.
- Pitcher, G. 1970. The awfulness of pain. *Journal of Philosophy* 48.
- Pitcher, G. 1970. Pain perception. *Philosophical Review* 74:368-93.
- Puccetti, R. 1975. Is pain necessary? *Philosophy* 50:259-69.
- Quinn, W. 1968. Pleasure -- disposition or episode? *Philosophy and Phenomenological Research* 28:578-86.
- Rachels, S. 2000. Is unpleasantness intrinsic to unpleasant experiences? *Philosophical Studies* 99:187-210.
- Rachlin, H. 1985. Pain and behavior. *Behavioral and Brain Sciences* 8:43-83.
- Resnik, D. 2000. Pain as a folk psychological concept: A clinical perspective. *Brain and Mind* 1:193-207.
- Sufka, K.J. & Lynch, M.P. 2000. Sensations and pain processes. *Philosophical Psychology* 13:299-311.
- Tye, M. 1995. A representational theory of pains and their phenomenal character. *Philosophical Perspectives* 9:223-39.
- Williams, B. 1959. Pleasure and belief. *Proceedings of the Aristotelian Society*.

5.2b Emotions -----

- Addis, L. 1995. The ontology of emotion. *Southern Journal of Philosophy* 33:261-78.
- Arregui, J.V. 1996. On the intentionality of moods: Phenomenology and linguistic analysis. *American Catholic Philosophical Quarterly* 70:397-411.
- Bedford, E. 1957. Emotions. *Proceedings of the Aristotelian Society* 57:281-304.
- Ben-Ze'ev, A. 1987. The nature of emotions. *Philosophical Studies* 52:393-409.
- Ben-Ze'ev, A. 1990. Describing the emotions. *Philosophical Psychology* 3:305-17.
- Ben-Ze'ev, A. 1992. Emotional and moral evaluations. *Metaphilosophy* 23:214-29.
- Charland, L.C. 1995. Feeling and representing: Computational theory and the modularity of affect. *Synthese* 105:273-301.
- D'Arms, J. & Jacobson, D. 2000. The moralistic fallacy: On the "appropriateness" of emotions. *Philosophy and Phenomenological Research* 61:65-90.
- Davis, W. 1981. A theory of happiness. *American Philosophical Quarterly* 18:111-20.
- DeLancey, C. 1997. Emotion and the computational theory of mind. In (S. O'Neillain, P. McKevitt, & E. MacAogain, eds) *Two Sciences of Mind*. John Benjamins.
- DeLancey, C. 1998. Real emotions. *Philosophical Psychology* 11:467-487.

- de Sousa, R. 1979. The rationality of emotions. Dialogue.
- de Sousa, R. 1987. *The Rationality of Emotion*. MIT Press.
- Deigh, J. 1994. Cognitivism in the theory of emotions. *Ethics* 104:824-54.
- Goldie, P. 2000. *The Emotions: A Philosophical Exploration*. Oxford University Press.
- Gordon, R.M. 1974. The aboutness of emotions. *American Philosophical Quarterly* 27:11-36.
- Gordon, R.M. 1986. The passivity of emotions. *Philosophical Review* 95:339-60.
- Gordon, R.M. 1987. *The Structure of Emotions: Investigations in Cognitive Philosophy*. Cambridge University Press.
- Green, O. 1992. *The Emotions: A Philosophical Theory*. Kluwer.
- Greenspan, P.S. 1988. *Emotions and Reasons: An Enquiry into Emotional Justification*. Routledge.
- Griffiths, P. 1989. Folk, functional and neurochemical aspects of mood. *Philosophical Psychology* 2:17-32.
- Haybron, D.M. 2001. Happiness and pleasure. *Philosophy and Phenomenological Research* 62:501-528.
- Helm, B.W. 1994. *The Significance of Emotions*. *American Philosophical Quarterly* 31:319-31.
- Irani, K.S. & Myers, G. 1983. *Emotion: Philosophical Studies*. Haven.
- Letwin, O. 1987. *Ethics, Emotion, and the Unity of the Self*. Croom Helm.
- Lormand, E. 1985. Toward a theory of moods. *Philosophical Studies* 47:385-407.
- Lyons, W. 1978. Emotions and behavior. *Philosophy and Phenomenological Research*.
- Marks, J. 1982. A theory of emotion. *Philosophical Studies* 42:227-42.
- McCullagh, C.B. 1990. The rationality of emotions and of emotional behavior. *Australasian Journal of Philosophy* 68:44-58.
- Morreal, J. 1983. Humor and emotion. *American Philosophical Quarterly* 20:297-304.
- Nash, R.A. 1989. Cognitive theories of emotion. *Nous* 23:481-504.
- Neu, J. 1971. *Emotion, Thought, and Therapy*. Cambridge University Press.
- Neu, J. 2000. *A Tear is an Intellectual Thing: The Meanings of Emotion*. Oxford University Press.
- Pugmire, D. 1994. Real emotion. *Philosophy and Phenomenological Research* 54:105-22.
- Rey, G. 1980. Functionalism and the emotions. In (A. Rorty, ed), *Explaining Emotions*. University of California Press.
- Roberts, R.C. 1995. Feeling one's emotions and knowing oneself. *Philosophical Studies* 77:319-38.

- Rorty, A.O. 1978. Explaining emotions. *Journal of Philosophy*.
- Rorty, A.O. (ed) 1980. *Explaining Emotions*. University of California Press.
- Rosenthal, D.M. 1983. Emotions and the self. In (K. Irani & G. Myers) *Emotion: Philosophical Studies*. Haven.
- Sizer, L. 2000. Towards a computational theory of mood. *British Journal for the Philosophy of Science* 51:743-770.
- Solomon, R.C. 1973. Emotion and choice. *Review of Metaphysics* 17:20-41.
- Taylor, G. 1975. Justifying the emotions. *Mind*.
- Thalberg, I. 1964. Emotion and thought. *American Philosophical Quarterly*.
- Wilkinson, S. 2000. Is 'normal grief' a mental disorder? *Philosophical Quarterly* 50:289-305.
- Wilson, J.R.S. 1972. *Emotion and Object*. Cambridge University Press.
- Wollheim, R. 1999. *On the Emotions*. Yale University Press.

5.2c Dreams [see also 6.2l]

- Ayer, A. 1960. Professor Malcolm on dreams. *Journal of Philosophy*.
- Chappell, V.C. 1963. The concept of dreaming. *Philosophical Quarterly* 13:193-213.
- Chihara, C. 1965. What dreams are made of. *Theoria* 31:145-58.
- Curley, E.M. 1975. Dreaming and conceptual revision. *Australasian Journal of Philosophy* 53:119-41.
- Dennett, D.C. 1976. Are dreams experiences? *Philosophical Review* 73:151-71. Reprinted in *Brainstorms* (MIT Press, 1978).
- Dunlop, C.E.M. (ed) 1977. *Philosophical Essays on Dreaming*. Cornell University Press.
- Emmett, K. 1978. Oneiric experiences. *Philosophical Studies* 34:445-50.
- Flanagan, O. 1995. Deconstructing dreams: The spandrels of sleep. *Journal of Philosophy* 92:5-27.
- Flanagan, O. 1996. Self-expression in sleep: Neuroscience and dreams. In *Self-Expressions*. Oxford University Press.
- Flanagan, O. 2000. *Dreaming Souls: Sleep, Dreams, and the Evolution of the Conscious Mind*. Oxford University Press.
- Hunter, J. 1971. Some questions about dreaming. *Mind* 80:70-92.
- Hunter, J. 1983. The difference between dreaming and being awake. *Mind* 92:80-93.
- Landesman, C. 1964. Dreams: Two types of explanation. *Philosophical Studies* 15:17-23.
- Malcolm, N. 1962. *Dreaming*. Routledge and Kegan Paul.
- Mannison, D.S. 1975. Dreaming an impossible dream. *Canadian Journal of Philosophy* 4:663-75.

Matthews, G.B. 1981. On being immoral in a dream. *Philosophy* 56:47-64.

Putnam, H. 1962. Dreaming and 'depth grammar'. In (R. Butler, ed) *Analytical Philosophy: First Series*. Oxford University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Schroeder, S. 1997. The concept of dreaming: On three theses by Malcolm. *Philosophical Investigations* 20:15-38.

Seligman, M. & Yellen, A. 1987. What is a dream? *Behavior Research and Therapy* 25:1-24.

Shanon, B. 1983. Descartes' puzzle -- An organismic approach. *Cognition and Brain Theory* 6:185-95.

Siegler, F.A. 1967. Remembering dreams. *Philosophical Quarterly* 17:14-24.

5.2d Memory [see also 2.2f, 3.7]

Arcaya, J.M. 1989. Memory and temporality: A phenomenological alternative. *Philosophical Psychology* 2:101-110.

Baier, A. 1976. Mixing memory and desire. *American Philosophical Quarterly* 13:213-20.

Ben-Zeev, A. 1986. Two approaches to memory. *Philosophical Investigations* 9:288-301.

Bergson, H. 1991. *Matter and memory*. MIT Press.

Campbell, J. 1997. The realism of memory. In (R. Heck, ed) *Language, Thought, and Logic: Essays in Honour of Michael Dummett*. Oxford University Press.

Campbell, J. 1997. The structure of time in autobiographical memory. *European Journal of Philosophy* 5:105-17.

Cherniak, C. 1983. Rationality and the structure of human memory. *Synthese* 57:163-86.

Furlong, E.J. 1956. The empiricist theory of memory. *Mind* 65:542-47.

Furlong, E.J. 1951. *A Study in Memory: A Philosophical Essay*. Nelson.

Gennaro, R.J. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333-47.

Haight, D. & Haight, M. 1989. Time, memory, and self-remembering. *Journal of Speculative Philosophy* 3:1-11.

Holland, R.F. 1954. The empiricist theory of memory. *Mind* 63:464-86.

Judson, L. 1988. Russell on memory. *Proceedings of the Aristotelian Society* 88:65-82.

Kurtzman, H.S. 1983. Modern conceptions of memory. *Philosophy and Phenomenological Research* 44:1-20.

Malcolm, N. 1970. Memory and representation. *Nous* 4:59-71.

Martin, M.G.F. 1992. Perception, concepts, and memory. *Philosophical Review* 101:745-63.

Munsat, S. 1979. Memory and causality. In (D. Gustafson, ed) *_Body, Mind, and Method_*. Reidel.

Naylor, A. 1985. In defense of a nontraditional theory of memory. *Monist* 62:136-50.

Owens, D. 1996. A Lockean theory of memory experience. *Philosophy and Phenomenological Research* 56:319-32.

Perkins, R.K. 1973. Russell on memory. *Mind* 82:600-1.

Rakover, S. 1983. In defense of memory viewed as stored mental representation. *Behaviorism* 11:53-62.

Rosen, D.A. 1975. An argument for the logical notion of a memory trace. *Philosophy of Science* 42:1-10.

Rundle, B. 1986. Memory and causation. *Philosophical Investigations* 9:302-7.

Rychlak, J.F. 1996. Memory: A logical learning account. *Journal of Mind and Behavior* 17:229-50.

Sanders, J.T. 1985. Experience, memory, and intelligence. *Monist* 68:507-21.

Schechtman, M. 1994. The truth about memory. *Philosophical Psychology* 7:3-18.

Shope, R.K. 1973. Remembering, knowledge, and memory traces. *Philosophy and Phenomenological Research* 33:303-22.

Stern, D.G. 1991. Models of memory: Wittgenstein and cognitive science. *Philosophical Psychology* 4:203-18.

Sutton, J. 1998. *_Philosophy and Memory Traces: Descartes to Connectionism_*. Cambridge University Press.

Urmson, J.O. 1971. Memory and imagination. *Mind* 80:607.

Wilcox, S. & Katz, S. 1981. A direct realistic alternative to the traditional conception of memory. *Behaviorism* 9:227-40.

Zemach, E. 1983. Memory: What it is, and what is cannot possibly be. *Philosophy and Phenomenological Research* 44:31-44.

5.2e Color [see also 1.3a, 1.7a, 1.7d]

Armstrong, D.M. 1969. Colour realism and the argument from microscopes. In (R. Brown & C. Rollins, eds) *_Contemporary Philosophy in Australia_*. Humanities Press.

Averill, E.W. 1985. Color and the anthropocentric problem. *Journal of Philosophy* 82:281-303.

Averill, E.W. 1992. The relational nature of color. *Philosophical Review* 101:551-88.

Bigelow, J. Collins, J. & Pargetter, R. 1990. Colouring in the world. *Mind* 99:279-88.

Boghossian, P. & Velleman, J.D. 1989. Color as a secondary quality. *Mind* 98:81-103.

Boghossian, P. & Velleman, J.D. 1991. Physicalist theories of color.

Philosophical Review 100:67-106.

Broackes, J. 1992. The autonomy of colour. In (D. Charles & K. Lennon, ed) Reduction, Explanation, and Realism. Oxford University Press.

Broackes, J. 1997. The Nature of Colour. Routledge.

Byrne, A. & Hilbert, D.R. 1997. Readings on Color, Volume 1: The Philosophy of Color. MIT Press.

Byrne, A. & Hilbert, D.R. 1997. Readings on Color, Volume 2: The Science of Color. MIT Press.

Byrne, A. & Hilbert, D.R. 1997. Colors and reflectances. In (A. Byrne & D.R. Hilbert, eds) Readings on Color, Volume 1: The Philosophy of Color. MIT Press.

Campbell, J. 1993. A simple view of colour. In (J. Haldane & C. Wright, ed) Reality, Representation, and Projection. Oxford University Press.

Campbell, K. 1969. Colours. In (R. Brown & C. Rollins, eds) Contemporary Philosophy in Australia. Humanities Press.

Campbell, K. 1982. The implications of Land's theory of colour vision. In (L. Cohen, ed) Logic, Methodology, and Philosophy of Science, Vol. 6. North-Holland.

Campbell, K. 1993. David Armstrong and realism about colour. In (J. Bacon, K. Campbell, & L. Reinhardt, eds) Ontology, Causality, and Mind. Cambridge University Press.

Clark, A. 1996. True theories, false colors. *Philosophy of Science Supplement* 63:143-50.

Dedrick, D. 1995. Objectivism and the evolutionary value of color vision. *Dialogue* 34:35-44.

Dedrick, D. 1996. Can color be reduced to anything? *Philosophy of Science Supplement* 3:134-42.

Foti, V.M. 1990. The dimension of color. *International Studies in Philosophy* 22:13-28.

Gilbert, P. 1987. Westphal and Wittgenstein on white. *Mind* 76:399-403.

Gilbert, P. 1989. Reflections on white: A rejoinder to Westphal. *Mind* 98:423-6.

Gold, I. 1999. Dispositions and the central problem of color.

Gold, I. 1999. On Lewis on naming the colours. *Australasian Journal of Philosophy* 77:365-370.

Philosophical Studies 93:21-44.

Hall, R.J. 1996. The evolution of color vision without colors. *Philosophy of Science Supplement* 63:125-33.

Hardin, C.L. 1983. Colors, normal observers and standard conditions. *Journal of Philosophy* 80:806-13.

Hardin, C.L. 1984. A new look at color. *American Philosophical Quarterly* 21:125-33.

- Hardin, C.L. 1984. Are scientific objects colored? *Mind* 93:491-500.
- Hardin, C.L. 1985. The resemblances of colors. *Philosophical Studies* 48:35-47.
- Hardin, C.L. 1985. Frank talk about the colors of sense-data. *Australasian Journal of Philosophy* 63:485-93.
- Hardin, C.L. 1988. *_Color for Philosophers_*. Hackett.
- Hardin, C.L. 1988. Phenomenal colors and sorites. *Nous* 22:213-34.
- Hardin, C.L. 1989. Could white be green? *Mind* 390:285-8.
- Hardin, C.L. 1989. Idle colors and busy spectra. *Analysis* 49:47-8.
- Hardin, C.L. 1990. Color and illusion. In (W. Lycan, ed) *_Mind and Cognition_*. Blackwell.
- Hardin, C.L. 1993. van Brakel and the not-so-naked emperor. *British Journal for the Philosophy of Science* 44:137-50.
- Harvey, J. 1992. Challenging the obvious: The logic of color concepts. *Philosophia* 21:277-94.
- Harvey, J. 2000. Colour-dispositionalism and its recent critics. *Philosophy and Phenomenological Research* 61:137-156.
- Hazen, A.P. 1999. On naming the colours. *Australasian Journal of Philosophy* 77:224-231.
- Hilbert, D.R. 1987. *_Color and Color Perception: A Study in Anthropocentric Realism_*. CSLI Press.
- Hilbert, D.R. 1992. What is color vision? *Philosophical Studies* 68:351-70.
- Jackson, F. 1996. The primary quality view of color. *Philosophical Perspectives* 10:199-219.
- Jackson, F. 1998. Colour, disjunctions, programming. *Analysis* 58:86-88.
- Jackson, F. & Pargetter, R. 1987. An objectivist's guide to subjectivism about color. *Revue Internationale de Philosophie* 41:127-v41.
- Jacovides, M. 2000. Cambridge changes of color. *Pacific Philosophical Quarterly* 81:142-164.
- Johnston, M. 1992. How to speak of the colors. *Philosophical Studies* 68:221-263.
- Kliewer, G. 1998. Neutral color concepts. *Philosophical Studies* 91:21-41.
- Kraut, R. 1992. The objectivity of color and the color of objectivity. *Philosophical Studies* 3:265-87.
- Langsam, H. 2000. Why colours do look like dispositions. *Philosophical Quarterly* 50:68-75.
- Landesman, C. 1989. *_Color and Consciousness: An Essay in Metaphysics_*. Temple University Press.
- Levin, J. 2000. Dispositional theories of color and the claims of common sense. *Philosophical Studies* 100:151-174.
- Lewis, D. 1997. Naming the colours. *Australasian Journal of Philosophy*

75:325-42.

Maund, J.B. 1981. Colour: A case for conceptual fission. *Australasian Journal of Philosophy* 59:308-22.

Maund, J.B. 1991. The nature of color. *History of Philosophy Quarterly* 8:253-63.

Maund, J.B. 1995. *_Colours: Their Nature and Representation_*. Cambridge University Press.

McFarland, D. & Miller, A. 1998. Jackson on colour as a primary quality. *Analysis* 58:76-85.

McFarland, D. & Miller, A. 2000. Disjunctions, programming and the Australian view of colour. *Analysis* 60:209-212.

McGilvray, J.A. 1983. To color. *Synthese* 54:37-70.

McGilvray, J.A. 1994. Constant colors in the head. *Synthese* 100:197-239.

McGinn, C. 1996. Another look at color. *Journal of Philosophy* 93:537-53.

McGinn, M. 1991. Westphal on the physical basis of color incompatibility. *Analysis* 4:218-22.

McGinn, M. 1991. On two recent accounts of color. *Philosophical Quarterly* 41:316-24.

Miller, A. 2001. The missing-explanation argument revisited. *Analysis* 61:76-86.

Montgomery, R. 1996. The indeterminacy of color vision. *Synthese* 106:167-203.

Nida-Rumelin, M. 1997. The character of color predicates: A phenomenalist view. In (M. Anduscius, A. Newen, & W. Kunne, eds) *_Direct Reference, Indexicality, and Propositional Attitudes_*. CSLI Press.

Ross, P.W. 1999. The appearance and nature of color. *Southern Journal of Philosophy* 37:227-252.

Ross, P. 2000. The relativity of color. *Synthese* 123:105-130.

Smart, J.J.C. 1975. On some criticisms of a physicalist theory of colors. In (C. Cheng, ed) *_Philosophical Aspects of the Mind-Body Problem_*. University Press of Hawaii.

Smart, J.J.C. 1995. 'Looks red' and dangerous talk. *Philosophy* 70_545-54.

Smith, M.A. Color, transparency, mind-independence. In (J. Haldane & C. Wright, ed) *_Reality, Representation, and Projection_*. Oxford University Press.

Smith, P. 1987. Subjectivity and colour vision. *Proceedings of the Aristotelian Society* 61:245-81.

Spohn, W. 1997. The character of color predicates: A materialist view. In (M. Anduscius, A. Newen, & W. Kunne, eds) *_Direct Reference, Indexicality, and Propositional Attitudes_*. CSLI Press.

Strawson, G. 1989. Red and 'red'. *Synthese* 78:193-232.

Stroud, B. 2000. *_The Quest for Reality: Subjectivism and the Metaphysics of Colour_*. Oxford University Press.

Stroud-Drinkwater, C. 1994. The naive theory of color. *Philosophy and Phenomenological Research* 54:345-54.

Thompson, E., Palacios, A., & Varela, F.J. 1992. Ways of coloring. *Behavioral and Brain Sciences*.

Thompson, E. 1995. Colour vision, evolution, and perceptual content. *Synthese* 104:1-32.

Thompson, E. 1995. *_Colour Vision_*. Routledge.

Tolliver, J.T. 1996. Interior colors. *Philosophical Topics* 22:411-41.

van Brakel, J. 1993. The plasticity of categories: The case of color. *British Journal for the Philosophy of Science* 44:103-135.

Watkins, M. 1999. Do animals see colors? An anthropocentrist's guide to animals, the color blind, and far away places. *Philosophical Studies* 94:189-209.

Westphal, J. 1982. Brown: Remarks on color. *Inquiry* 25:417-33.

Westphal, J. 1986. White. *Mind* 95:310-28.

Westphal, J. 1989. Black. *Mind* 98:585-9.

Westphal, J. 1991. *_Colour: A Philosophical Introduction_*. Blackwell.

Whitmyer, VG. 1999. Ecological color. *Philosophical Psychology* 12:197-214.

Wittgenstein, L. 1977. *_Remarks on Colour_*. University of California Press.

5.3 Philosophy of Psychology, General

5.3a Psychological Laws [see also 3.5d]

Antony, L. 1995. Law and order in psychology. *Philosophical Perspectives* 9:429-46.

Braithwaite, M. 1949. Causal laws in psychology. *Aristotelian Society Supplement* 23:45-60.

Fodor, J.A. 1991. You can fool some of the people all of the time, everything else being equal: Hedged laws and psychological explanation. *Mind* 100:19-34.
Ceteris paribus means that every realizing state has completing conditions.
Even absolute exceptions are OK, as long as they're not across-the-board.

Fodor, J.A. 1989. Making mind matter more. *Philosophical Topics* 17:59-79.
Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).
Non-strict psychological laws are compatible with the (nomologically sufficient) causal responsibility of mental properties. So there's no need for epiphobia. With comments on the relation between laws and mechanisms.

Horgan, T. & Tienson, J. 1990. Soft laws. *Midwest Studies in Philosophy* 15.
Argues that any laws in intentional psychology have ineliminable same-level exceptions; the Kuhnian crisis in cognitive science gives evidence for this.
But ceteris paribus laws provide perfectly good theoretical explanation.

Lycan, W.G. 1981. Psychological laws. *Philosophical Topics* 12:9-38.
A functionalist defense against anomalous monism. Psychofunctional laws and psychological laws, though not psychophysical laws, may exist. Rebutting

arguments from rationality, indeterminism, intensionality, etc.

Mace, C.A. 1949. Causal laws in psychology. Aristotelian Society Supplement 23:61-68.

Marcello, G. 2000. Horgan and Tienson on ceteris paribus laws. Philosophy of Science 67:301-315.

Mott, P. 1992. Fodor and ceteris paribus laws. Mind 101:335-46.

Pietroski, P. & Rey, G. 1995. When other things aren't equal: Saving ceteris paribus laws from vacuity. British Journal for the Philosophy of Science 46:81-110.

Schiffer, S. 1991. Ceteris paribus laws. Mind 100:1-17.

There are no ceteris paribus laws, as there's no satisfactory way to cash the "unless" cause. But psychology doesn't need laws, anyway.

Silverberg, A. 1996. Psychological laws and nonmonotonic logic. Erkenntnis 44:199-224.

Warfield, T.A. 1993. Folk-psychological ceteris-paribus laws. Philosophical Studies 71:99-112.

5.3b Psychology and Neuroscience [see also 6.1i]

Bechtel, W. 1983. A bridge between cognitive science and neuroscience: The functional architecture of mind. Philosophical Studies 44:319-30.

Arguing for the notion of functional architecture as a bridge whereby neural components can be components of cognitive processes.

Bub, J. 1994. Testing models of cognition through the analysis of brain-damaged patients. British Journal for the Philosophy of Science 45:837-55.

Butler, K. 1994. Neural constraints in cognitive science. Minds and Machines 4:129-62.

Cherniak, C. 1991. Meta-neuroanatomy: The myth of the unbounded mind/brain. In (E. Agazzi, ed) _Philosophy and the Origin and Evolution of the Universe_.

Cherniak, C. 1994. Philosophy and computational neuroanatomy. Philosophical Studies 73:89-107.

Argues that we can understand the brain under the hypothesis that it is optimized to "save wire", due to bounded resources: organization predicts placement. With remarks on the relation between cognitive and neural levels.

Churchland, P.M. 1986. Some reductive strategies in cognitive neurobiology. Mind 95:279-309. Reprinted in _A Neurocomputational Perspective_ (MIT Press, 1989).

Some cute examples of neurophysiological reductions using state-spaces.

Churchland, P.M. 1995. _The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain_. MIT Press.

Churchland, P.S. 1980. A perspective on mind-brain research. Journal of Philosophy 77:185-207.

The brain can tell us a lot about the mind. With examples.

Churchland, P.S. 1982. Mind-brain reduction: New light from philosophy of science. Neuroscience 7:1041-7.

Churchland, P.S. 1986. *Neurophilosophy: Toward A Unified Science of the Mind-Brain*. MIT Press.

All about neuroscience, philosophy and prospects for their interaction.

Churchland, P.S. & Sejnowski, T. 1989. Neural representation and neural computation. In (L. Nadel, ed) *Neural Connections, Mental Computations*. MIT Press.

About how neuroscience and connectionism affect our conception of mind.

Churchland, P.S. 1987. Epistemology in the age of neuroscience. *Journal of Philosophy* 84:546-53.

On paradigm shifts, biology, evolution, connectionism, etc.

Clark, A. 1980. *Psychological Models and Neural Mechanisms: An Examination of Reductionism in Psychology*. Oxford University Press.

Glymour, C. 1994. On the methods of cognitive neuropsychology. *British Journal for the Philosophy of Science* 45:815-35.

Hardcastle, V.G. 1992. Reduction, explanatory extension, and the mind/brain sciences. *Philosophy of Science* 59:408-28.

The relationship between psychology and neuroscience is best characterized not by reduction but by explanatory extension, where each field is enriched by the other. With a number of examples from recent empirical work.

Hatfield, G. 1988. Neurophilosophy meets psychology: Reduction, autonomy, and empirical constraints. *Cognitive Neuropsychology* 5:723-46.

Hatfield, G. 2000. The brain's 'new' science: Psychology, neurophysiology, and constraint. *Philosophy of Science* 67:388-404.

Klagge, J.C. 1989. Wittgenstein and neuroscience. *Synthese* 78:319-43.

Wittgenstein wouldn't have liked the Churchlands, as neuro might be chaos, and too much neuro might undermine our self-conception nihilistically.

Kobes, B. 1991. On a model for psycho-neural coevolution. *Behavior and Philosophy* 19:1-17.

Madell, G. 1986. Neurophilosophy: A principled skeptic's response. *Inquiry*.

Manier, E. 1986. Problems in the development of cognitive neuroscience: Effective communication between scientific domains. *Philosophy of Science Association* 1986, 1:183-97.

McCauley, R. 1986. Intertheoretic relations and the future of psychology. *Philosophy of Science* 53:179-99.

Incommensurable theories don't necessarily require elimination, if their relationship is synchronic/interlevel, rather than diachronic/intralevel.

Mucciolo, L. 1974. The identity thesis and neuropsychology. *Nous* 8:327-42.

Argues contra Fodor and Block that neurological equipotentiality doesn't refute type materialism. Mental states may not be anatomically defined neural states, but they may be more abstract neural holograms.

Mundale, J. & Bechtel, W. 1996. Integrating neuroscience, psychology, and evolutionary biology through a teleological conception of function. *Minds and Machines* 6:481-505.

Ravenscroft, I. 1998. Neuroscience and the mind. *Mind and Language* 13:132-137.

Rockwell, W.T. 1994. On what the mind is identical with. *Philosophical Psychology* 7:307-23.

Argues that the mind is not identical with the brain -- at the very least, it's the central nervous system, and perhaps more. "Brain" does not denote a natural kind in neurophysiology.

Smith, A. 1986. Brain-mind philosophy. *Inquiry* 29:203-15.

Skarda, S. 1986. Explaining behavior: Bringing the brain back in. *Inquiry* 29:187-201.

Stoljar, D. & Gold, S. 1998. On biological and cognitive neuroscience. *Mind and Language* 13:110-31.

Stone, T. & Davies, M. 1993. Cognitive neuropsychology and the philosophy of mind. *British Journal for the Philosophy of Science* 44:589-622.

Stufflebeam, R.S. & Bechtel, W. 1997. PET: Exploring the myth and the method. *Philosophy of Science* 64:95-106.

van Orden, G.C. 1997. Functional neuroimages fail to discover pieces of mind in the parts of the brain. *Philosophy of Science Supplement* 64:85-94.

von Eckardt, B. 1984. Cognitive psychology and principled skepticism. *Journal of Philosophy* 81:67-88.

Cognitive psychology can transmogrify itself, who needs neuroscience?

5.3c Explanation in Cognitive Science

Cummins, R. 1982. The internal manual model of psychological explanation. *Cognition and Brain Theory* 5:257-68.

Cummins, R. 1983. *The Nature of Psychological Explanation*. MIT Press.
Psychological explanation is typically via functional analysis, not causal subsumption. On interpretation, computation, and an analysis of cognition and intentionality. With remarks on Dretske, Searle, Titchener, Hull, Freud.

Fodor, J.A. 1968. *Psychological Explanation*. Random House.

Fodor, J.A. 1968. The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy* 65:627-40. Reprinted in *RePresentations* (MIT Press, 1980).

Franks, B. 1995. On explanation in cognitive science: Competence, idealization, and the failure of the classical cascade. *British Journal for the Philosophy of Science* 46:475-502.

Gilman, D. 1993. Optimization and simplicity: Marr's theory of vision and biological explanation. *Synthese* 107:293-323.
Contra Kitcher 1988, much of Marr's theory doesn't depend on optimization; in any case, optimization isn't so bad. With remarks on interdisciplinarity.

Heil, J. 1986. Formalism and psychological explanation. *Journal of Mind and Behavior* 7:1-10.

On the tension between formal explanation and representational explanation.

Kim, J. 1989. Mechanism, purpose, and explanatory exclusion. **Philosophical Perspectives** 3:77-108. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Discusses the principle: there cannot be two independent explanations of the same phenomena. With application to purposive explanation of behavior, theory reduction, and eliminativism, and a discussion of explanatory realism.

Kim, J. 1990. Explanatory exclusion and the problem of mental causation. In

(E. Villanueva, ed) *_Information, Semantics, and Epistemology_*. Blackwell.
On the problems posed by explanatory exclusion, and possible solutions. With
focus on the problems as they arise for Dretske's and Davidson's theories.

Knight, D. 1997. A poetics of psychological explanation. *Metaphilosophy*
28:63-80.

Millikan, R.G. 1993. Explanation in biopsychology. In (J. Heil & A. Mele,
eds) *_Mental Causation_*. Oxford University Press.

Montgomery, R. 1995. Explanation and evaluation in cognitive science.
Philosophy of Science 62:261-82.

Montgomery, R. 1998. Grades of explanation in cognitive science. *Synthese*
114:463-495.

Morris, M. 1986. Causes of behavior. *Philosophical Quarterly* 36:123-44.

Moser, P. 1994. Naturalism and psychological explanation. *Philosophical*
Psychology 7:63-84.

Owens, J. 1998. Psychological explanation and causal deviancy. *Synthese*
115:143-169.

Sober, E. 1978. Psychologism. *Journal for the Theory of Social Behavior*
8:165-91.

5.3d Philosophy of Cognitive Science, Misc

Bealer, G. 1987. The boundary between philosophy and cognitive science.
Journal of Philosophy 86:553-55.

Philosophy is autonomous: empirical considerations can't affect it.

Bogdan, R. 2000. *_Minding Minds: Evolving a Reflexive Mind by Interpreting*
Others_. MIT Press.

Fetzer, J.H. 1991. *_Philosophy and Cognitive Science_*. Paragon House.

Flanagan, O.J. 1984. *_The Science of the Mind_*. MIT Press.

Hardcastle, V. 1996. *_How to Build a Theory in Cognitive Science_*. SUNY
Press.

Harnad, S. 1982. Neoconstructivism: A unifying constraint for cognitive
science. In (T. Simon & R. Scholes, eds) *_Language, Mind, and Brain_*.
Lawrence Erlbaum.

Haugeland, J. 1978. The nature and plausibility of cognitivism. *Behavioral*
and Brain Sciences 1:215-26.

Hooker, C.A. 1975. The information-processing approach to the brain-mind and
its philosophical ramifications. *Philosophy and Phenomenological Research*
36:1-15.

Keely, B. 2000. Neuroethology and the philosophy of cognitive science.
Philosophy of Science 67:404-418.

Kukla, A. 1989. Non-empirical issues in psychology. *American Psychologist*
44:485-94.

On the role of non-empirical advances in psychology: e.g. in theory
construction, coherence analysis, conceptual innovation, with the aid of
logically necessary truths and the contingent/pragmatic a priori.

Lloyd, D. 1989. *_Simple Minds_*. MIT Press.

O'Neill, S. 1995. *_The Search for Mind: A New Foundation for Cognitive Science_*. Ablex.

O'Neill, S., McKevitt, P. & MacAogain, E. (eds) 1997. *_Two Sciences of Mind_*. John Benjamins.

Pickering, M. & Chater, N. 1995. Why cognitive science is not formalized folk psychology. *Minds and Machines* 5.

Preston, B. 1994. Behaviorism and mentalism: Is there a third alternative? *Synthese* 100:167-96.

van Gelder, T. 1998. The roles of philosophy in cognitive science. *Philosophical Psychology* 11:117-36.

von Eckardt, B. 1993. *_What is Cognitive Science?_* MIT Press.

5.4 Philosophy of Mind, General

Armstrong, D. 1999. *_The Mind-Body Problem: An Opinionated Introduction_*. Westview Press.

Baker, L.R. 1989. Recent work in the philosophy of mind. *Philosophical Books* 30:1-9.
A general overview.

Bealer, G. 1986. The logical status of mind. *Midwest Studies in Philosophy* 10.

Bechtel, W. 1988. *_Philosophy of Mind: An Overview for Cognitive Science_*. Lawrence Erlbaum.

Braddon-Mitchell, D. & Jackson, F. 1997. *_Philosophy of Mind and Cognition_*. Blackwell.

Burge, T. 1992. Philosophy of language and mind: 1950-1990. *Philosophical Review* 100:3-52.
An overview of the last 40 years of the philosophy of language and the philosophy of mind, covering many issues and trends.

Carruthers, P. 1986. *_Introducing Persons: Theories and Arguments in the Philosophy of Mind_*. SUNY Press.

Churchland, P.M. 1984. *_Matter and Consciousness_*. MIT Press.

Crane, T. 1995. *_The Mechanical Mind_*. Penguin.

Dennett, D.C. 1978. Current issues in the philosophy of mind. *American Philosophical Quarterly* 15:249-261.
An overview of everything, circa 1978: logical behaviorism, functionalism, the identity theory, qualia, meaning, and so on, with bibliography.

Graham, G. 1993. *_Philosophy of Mind: An Introduction_*. Blackwell.

Guttenplan, S. 2000. *_Mind's Landscape: An Introduction to the Philosophy of Mind_*. Blackwell Publishers.

Haldane, J.J. 1994. Analytical philosophy and the nature of mind: Time for another rebirth? In (R. Warner & T. Szubka, eds) *_The Mind-Body Problem: A Guide to the Current Debate_*. Blackwell.

Haldane, J. 2000. The state and fate of contemporary philosophy of mind. *American Philosophical Quarterly* 37:301-21.

Hannay, B. 1994. *Subjectivity and Reduction: An Introduction to the Mind-Body Problem*. Westview Press.

Harman, G. 1989. Some philosophical issues in cognitive science. In (M. Posner, ed) *Foundations of Cognitive Science*. MIT Press.

Kim, J. 1996. *Philosophy of Mind*. Westview Press.

Lowe, E.J. 2000. *An Introduction to the Philosophy of Mind*. Cambridge University Press.

McGinn, C. 1982. *The Character of Mind*. Oxford University Press.

Phillips, H. 1995. *Vicissitudes of the I: An Introduction to the Philosophy of Mind*. Prentice-Hall.

Quine, W.V. 1985. States of mind. *Journal of Philosophy* 82:5-8.

Rey, G. 1997. *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Blackwell.

Rorty, R. 1982. Contemporary philosophy of mind. *Synthese* 53:323-48.
In praise of the "Ryle-Dennett" tradition, and the elimination of dualism from the philosophy of mind.

Rorty, R. 1993. Consciousness, intentionality, and pragmatism. In (S. Christensen & D. Turner, eds) *Folk Psychology and the Philosophy of Mind*. Lawrence Erlbaum.
A pragmatist perspective on the recent history of the philosophy of mind, focusing on consciousness, intentionality, and mental representation, and on debates between Fodor, Dennett, Searle, Putnam, and Davidson.

Shaffer, J.A. 1964. *Philosophy of Mind*. Prentice-Hall.

Smith, P. & Jones, O. 1986. *The Philosophy of Mind: An Introduction*. Cambridge University Press.

Sprague, E. 1999. *Persons and their Minds: A Philosophical Investigation*. Westview Press.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 6: Consciousness in Science [1599]

=====

- 6.1 Consciousness and Neuroscience [460]
 - 6.1a Neural Correlates of Consciousness [74]
 - 6.1b Neural Correlates of Visual Consciousness [28]
 - 6.1c Blindsight [50]
 - 6.1d Other Neuropsychological Disorders [71]
 - 6.1e Cerebral Hemispheres and Consciousness [45]
 - 6.1f Neural Timing and Consciousness (Libet) [19]
 - 6.1g Neural Synchrony and Binding [42]
 - 6.1h Consciousness and Anesthesia [44]
 - 6.1i Foundational Issues [31]
 - 6.1j Consciousness and Neuroscience, Misc [56]
- 6.2 Consciousness and Psychology [787]
 - 6.2a Cognitive Models of Consciousness [60]
 - 6.2b Unconscious Perception [73]
 - 6.2c Unconscious Processes [45]
 - 6.2d Visual Consciousness [37]
 - 6.2e Consciousness and Attention [35]
 - 6.2f Consciousness and Memory [84]
 - 6.2g Consciousness and Learning [33]
 - 6.2h Consciousness and Metacognition [22]
 - 6.2i Consciousness and Control [38]
 - 6.2j Consciousness and Imagery [24]
 - 6.2k Consciousness and Emotion [31]
 - 6.2l Consciousness, Sleep, and Dreaming [34]
 - 6.2m Consciousness in Psychoanalysis [57]
 - 6.2n Consciousness and Time [35]
 - 6.2o Self-Consciousness [46]
 - 6.2p Development of Consciousness [33]
 - 6.2q The Stream of Consciousness [29]
 - 6.2r Foundational Issues [19]
 - 6.2s Consciousness and Psychology, Misc [52]
- 6.3 Consciousness and Physics [77]
 - 6.3a Interpretation of Quantum Mechanics [27]
 - 6.3b Quantum Mechanisms of Consciousness [30]
 - 6.3c Consciousness and Physics, Misc [20]
- 6.4 Consciousness and Science, Misc [275]
 - 6.4a Evolution of Consciousness [31]
 - 6.4b Consciousness and Language [23]
 - 6.4c Animal Consciousness [33]
 - 6.4d Animal Self-Consciousness [22]
 - 6.4e Altered States of Consciousness [27]
 - 6.4f Parapsychology and Consciousness [54]
 - 6.4g Phenomenology [34]
 - 6.4h Foundations [27]
 - 6.4i Miscellaneous [24]

6.1 Consciousness and Neuroscience

6.1a Neural Correlates of Consciousness

Alkire, M.T., Haier, R.J., & James, H.F. 1998. Toward the neurobiology of consciousness: Using brain imaging and anesthesia to investigate the anatomy of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Baars, B.J. & Newman, J. 1994. A neurobiological interpretation of the Global

- Workspace theory of consciousness. In (A. Revonsuo & M. Kamppinen, eds) *_Consciousness in Philosophy and Cognitive Neuroscience_*. Lawrence Erlbaum.
- Baars, B.J. 1995. Surprisingly small subcortical structures are needed for the state of waking consciousness, while cortical projection areas seem to provide perceptual contents of consciousness. *Consciousness and Cognition* 4:159-62.
- Baars, B.J., Newman, J. & Taylor, J.G. 1998. Neuronal mechanisms of consciousness: A relational global workspace approach. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.
- Bogen, J.E. 1995. On the neurophysiology of consciousness, part I: An overview. *Consciousness and Cognition* 4:52-62.
- Bogen, J.E. 1995. On the neurophysiology of consciousness, part II: Constraining the semantic problem. *Consciousness and Cognition* 4:137-58.
- Bogen, J.E. 1997. Some neurophysiologic aspects of consciousness. *Seminars in Neurology* 17:95-103.
- Bogen, J.E. 1998. Locating the subjectivity pump: The thalamic intralaminar nuclei. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.
- Bremer, F. 1966. Neurophysiological correlates of mental unity. In (J. Eccles, ed) *_Brain and Conscious Experience_*. Springer.
- Buser, P.A. & Rougeul-Buser, A. (eds) 1978. *Cerebral correlates of conscious experience*. Elsevier.
- Coenen, A.M.L. 1998. Neuronal phenomena associated with vigilance and consciousness: From cellular mechanisms to electroencephalographic patterns. *Consciousness and Cognition* 7:42-53.
- Crick, F. 1984. Functions of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences USA* 81:4586-93.
- Crick, F. & Koch, C. 1990. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2:263-275.
- Crick, F. & Koch, C. 1998. *Consciousness and neuroscience*. Cerebral Cortex.
- Crick, F. & Koch, C. 2000. The unconscious homunculus. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Damasio, A. 2000. A neurobiology for consciousness. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.
- Dimond, S. 1976. Brain circuits for consciousness. *Brain, Behavior, and Evolution* 13:376-95.
- Duzel E., Yonelinas A.P., Mangun G.R., Heinze H.J., & Tulving E. 1997. Event-related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences of the United States of America* 94:5973-8.
- Edelman, G. & Tononi, G. 2000. Reentry and the dynamic core: Neural correlates of conscious experience. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Ellis, R. 2000. Efferent brain processes and the enactive approach to consciousness. *Journal Of Consciousness Studies* 7:40-50.

Flohr, H. 1990. Brain processes and phenomenal consciousness: A new and specific hypothesis. *Theory and Psychology* 1:245-62.

Flohr, H. 1992. Qualia and brain processes. In (A. Beckermann, H. Flohr, & J. Kim, eds) *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.

Flohr, H. 1995. Sensations and brain processes. *Behavioral Brain Research* 71:157-61.

Gallese, V. 2000. The acting subject: Toward the neural basis of social cognition. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Gazzaniga, M. 1993. Brain mechanisms and conscious experience. In *_Experimental and Theoretical Studies of Consciousness_* (Ciba Foundation Symposium 174). Wiley.

Gazzaniga, M. 1998. Brain and conscious experience. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Gray, J.A. 1995. The contents of consciousness: A neuropsychological conjecture. *Behavioral and Brain Sciences* 18:659-76.

Greenfield, S. 1997. How might the brain generate consciousness? *Communication and Cognition* 30:285-300.

Greenfield, S. 1998. A rosetta stone for mind and brain? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Grossenbacher, P.G. (ed) 1997. *_Finding Consciousness in the Brain: A Neurocognitive Approach_*. John Benjamins.

Goldman-Rakic, P.S. 1988. The prefrontal contribution to working memory and conscious experience. In (O. Creutzfeld & J. Eccles, eds) *_The Brain and Conscious Experience_*. Pontifical Academy.

Hobson, J.A. 1994. *_The Chemistry of Conscious States_*. Basic Books.

Hobson, J.A. 1997. Consciousness as a state-dependent phenomenon. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Jasper, H. 1998. Sensory information and conscious experience. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

John, E.R., Easton, P. & Isenhardt, R. 1997. Consciousness and cognition may be mediated by multiple independent coherent ensembles. *Consciousness and Cognition* 6:3-39.

Jones, B.E. 1998. The neural basis of consciousness across the sleep-waking cycle. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Jung, R. 1954. Correlation of bioelectrical and autonomic phenomena with alterations of consciousness and arousal in man. In (J. Delafresnaye, ed) *_Brain Mechanisms and Consciousness_*. Blackwell.

Kahn, D., Pace-Schott, E.F. & Hobson, J.A. 1997. Consciousness in waking and dreaming: The roles of neuronal oscillation and neuromodulation in determining similarities and differences. *Neuroscience* 78:13-38.

Kanwisher, N. 2001. Neural events and perceptual awareness. *Cognition* 79:89-113.

Kinsbourne, M. 1988. An integrated field theory of consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Kinsbourne, M. 1993. Integrated cortical field model of consciousness. In (Ciba Foundation) *_Experimental and Theoretical Studies of Consciousness_*. Wiley.

Kinsbourne, M. 1995. The intralaminar thalamic nuclei: Subjectivity pumps or attention-action co-ordinators? *Consciousness and Cognition* 4:167-71.

Kleitman, N. 1955. The role of the cerebral cortex in the development and maintenance of consciousness. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the Third Conference_*. Josiah Macy Foundation.

Koch, C. & Crick, F. 1994. Some further ideas regarding the neuronal basis of awareness. In (C. Koch & J. Davis, eds) *_Large-Scale Neuronal Theories of the Brain_*. MIT Press.

Koch, C. & Crick, F. 2000. Some thoughts on consciousness and neuroscience. In (M. Gazzaniga, ed) *_The New Cognitive Neurosciences: 2nd Edition_*. MIT Press.

Lehmann, D., Strik, W.K., Henggeler, B., Koenig, T. 1998. Brain electric microstates and momentary conscious mind states as building blocks of spontaneous thinking: I. Visual imagery and abstract thoughts. *International Journal of Psychophysiology* 29:1-11.

Libet, B. 1982. Brain stimulation in the study of neuronal functions for conscious sensory experiences. *Human Neurobiology* 1:235-42.

Libet, B. 1989. Conscious subjective experience vs. unconscious mental functions: A theory of the cerebral processes involved. In (R. Cotterill, ed) *_Models of Brain Function_*. Cambridge University Press.

Libet, B. 1996. Neural processes in the production of conscious experiences. In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.

Libet, B. 1998. Do the models offer testable proposals of brain functions for conscious experience? In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Magoun, H.W. 1954. The ascending reticular system and wakefulness. In (J. Delafresnaye, ed) *_Brain Mechanism and Consciousness_*. Blackwell.

Markowitsch H.J. 1995. Cerebral bases of consciousness: A historical view. *Neuropsychologia* 33:1181-1192.

Metzinger, T. 2000. *_Neural Correlates of Consciousness: Empirical and Conceptual Questions_*. MIT Press.

Newman, J.B. 1995. Thalamic contributions to attention and consciousness. *Consciousness and Cognition* 4:172-93.

Newman, J.B. 1997. Putting the puzzle together: Toward a general theory of the neural correlates of consciousness. *Journal of Consciousness Studies* 4:47-66, 4:100-121.

Newman, J.B. & Baars, B.J. 1993. A neural attentional model for access to consciousness: A global workspace perspective. *Concepts in Neuroscience*

4:255-90.

O'Keefe, J. 1985. Is consciousness the gateway to the hippocampal cognitive map? A speculative essay on the neural basis of mind. In (D. Oakley, ed) Brain and Mind. Methuen.

Orpwood R.D. 1994. A possible neural mechanism underlying consciousness based on the pattern processing capabilities of pyramidal neurons in the cerebral cortex. *Journal of Theoretical Biology* 169:403-18.

Parvizi, J. & Damasio, A. 2001. Consciousness and the brainstem. *Cognition* 79:135-59.

Penfield, W. 1937. The cerebral cortex and consciousness. In The Harvey Lectures. Reprinted in (R. Wilkins, ed) Neurosurgical Classics. Johnson Reprint Corporation, 1965.

Penfield, W. 1955. The permanent record of the stream of consciousness. *Acta Psychologica* 11:47-69.

Raichle, M. 2000. The neural correlates of consciousness: An analysis of cognitive skill learning. In (M. Gazzaniga, ed) The New Cognitive Neurosciences: 2nd Edition. MIT Press.

Rapcsak, S. & Kaszniak, A. 2000. Searching for the neural correlates of consciousness: Clues from face recognition research. *Brain & Cognition* 42:37-40.

Rudell, A.P. & Hua, J. 1996. The recognition potential and conscious awareness. *Electroencephalography and Clinical Neurophysiology* 98:309-318.

Sewards, T. & Sewards, M. 2000. The awareness of thirst: Proposed neural correlates. *Consciousness & Cognition* 9:463-487.

Smythies, J. 1997. The functional neuroanatomy of awareness: With a focus on the role of various anatomical systems in the control of intermodal attention. *Consciousness and Cognition* 6:455-81.

Sokolov, E.N. 1992. The neurophysiological mechanisms of consciousness. *Journal of Russian and East European Psychology* 30:6-12.

Strehler, B.L. 1991. Where is the self?: A neuroanatomical theory of consciousness. *Synapse* 7:44-91.

Stuss, D.T. 1991. Self, awareness, and the frontal lobes: A neuropsychological perspective. In (J. Strauss, ed) The Self: Interdisciplinary Approaches. Springer-Verlag.

Taylor, J.G. 1998. Cortical activity and the explanatory gap. *Consciousness and Cognition* 7:109-48.

Tononi, G. & Edelman, G. 1998. Consciousness and the integration of information in the brain. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) Consciousness: At the Frontiers of Neuroscience. Lippincott-Raven.

Umla, C. 2000. Conscious experience depends on multiple brain systems. *European Psychologist* 5:3-11.

Vanderwolf, C. 2000. Are neocortical gamma waves related to consciousness? *Brain Research* 855:217-224.

Verfaellie, M. & Keane, M.M. 1997. The neural basis of aware and unaware forms

of memory. *Seminars in Neurology* 17:153-61.

Woolf, N.J. 1997. A possible role for cholinergic neurons of the basal forebrain and pontomesencephalon in consciousness. *Consciousness and Cognition* 6:574-596.

6.1b Neural Correlates of Visual Consciousness (see also 6.1c)

Anderson, R.A. 1997. Neural mechanisms in visual motion perception in primates. *Neuron* 18:865-872.

Cowey, A. 1996. Visual awareness: Still at sea with seeing? *Current Biology* 6:45-47.

Crick, F. & Koch, C. 1995. Cortical areas in visual awareness. *Nature* 377:294-5.

Crick, F. & Koch, C. 1995. Are we aware of neural activity in primary visual cortex? *Nature* 375:121-23.

Farah, M.J., O'Reilly, R.C. & Vecera, S.P. 1997. The neural correlates of perceptual awareness: Evidence from covert recognition in prosopagnosia. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Farah, M. 2000. *_The Cognitive Neuroscience of Vision_*. Blackwell Publishers.

Ffytche, D. 2000. Imaging conscious vision. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Goodale, M.A. & Milner, A.D. 1992. Separate visual pathways for perception and action. *Trends in Neuroscience* 15:20-25.

Goodale, M. & Murphy, K. 2000. Space in the brain: Different neural substrates for allocentric and egocentric frames of reference. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Hubel, D.H. 1998. Recordings from the striate cortex in awake behaving animals. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Koch, C. 1995. Visual awareness and the thalamic intralaminar nuclei. *Consciousness and Cognition* 4:163-66.

Koch, C. 1996. Toward the neuronal substrate of visual consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Koch, C. 1998. The neuroanatomy of visual consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Koch, C. & Braun, J. 1996. Toward the neuronal correlate of visual awareness. *Current Opinion in Neurobiology* 6:158-64.

Leopold, D.A. & Logothetis, N.K. 1996. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* 379: 549-553.

Logothetis, N. & Schall, J. 1989. Neuronal correlates of subjective visual perception. *Science* 245:761-63.

Logothetis, N.K., Leopold, D.A. & Sheinberg, D.L. 1996. What is rivalling

during binocular rivalry? *Nature* 30(6575):621-624.

Logothetis, N.K. & Leopold, D.A. 1998. Single-neuron activity and visual perception. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Logothetis, N.K. 1999. Binocular rivalry: A window onto consciousness. *Scientific American*.

Lumer, E. 2000. Binocular rivalry and human visual awareness. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Marzi, C., Girelli, M., Miniussi, C., Smania, N., & Maravita, A. 2000. Electrophysiological correlates of conscious vision: Evidence from unilateral extinction. *Journal of Cognitive Neuroscience* 12:869-877.

Milner, A.D. 1995. Cerebral correlates of visual awareness. *Neuropsychologia* 33:1117-30.

Milner, A.D. 1998. Streams and consciousness: Visual awareness and the brain. *Trends in Cognitive Sciences* 2:25-30.

Milner, A.D. & Goodale, M.A. 1995. *The Visual Brain in Action*. Oxford University Press.

Sewards, T. & Sewards, M. 2000. Visual awareness due to neuronal activities in subcortical structures: A proposal. *Consciousness & Cognition* 9:86-116

Sheinberg, D.L. & Logothetis, N.K. 1997. The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences USA* 94:3408-3413.

Stoerig P. & Cowey A. 1995. Visual perception and phenomenal consciousness. *Behavioural Brain Research* 71:147-156.

Vanni, S., Revonsuo, A., Saarinen, J. & Hari, R. 1996. Visual awareness of objects correlates with activity of right occipital cortex. *Neuroreport* 8:183-186.

Zeki, S. & Bartels, A. 1999. Toward a theory of visual consciousness. *Consciousness and Cognition* 8:225-59.

6.1c Blindsight

Azzopardi, P., & Cowey, A. 1997. Is blindsight like normal, near-threshold vision? *Proceedings of the National Academy of Sciences USA* 94:14190.

Barbur, J.L., Watson, J.D.G., Frackowiak, R.D.G., & Zeki, S. 1993. Conscious visual perception without V1. *Brain* 116:1293-1302.

Braddick, O., Atkinson, J., Hood, B., Harkness, W. 1992. Possible blindsight in infants lacking one cerebral hemisphere. *Nature* 360:461-463.

Campion, J, Latto, R., & Smith, Y. 1983. Is blindsight an effect of scattered light, spared cortex, and near-threshold vision? *Behavioral and Brain Sciences* 6:423-86.

Carey, D.P., Goodale, M.A. & Sprowl, E.G. 1990. Blindsight in rodents: The use of a "high-level" distance cue in gerbils with lesions of primary visual cortex. *Behavioural Brain Research* 38:283-289.

Cowey, A. 1995. Blindsight in real sight. *Nature* 377:290-1.

Cowey, A. 1995. Blindsight in monkeys. *Nature* 373:247-9.

Cowey, A. & Stoerig, P. 1991. The neurobiology of blindsight. *Trends in Neurosciences* 14:140-5.

Cowey, A. & Stoerig, P. 1992. Reflections on blindsight. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Cowey, A. & Stoerig, P. 1997. Visual detection in monkeys with blindsight. *Neuropsychologia* 35:929-39.

Danckert, J. & Goodale, M. 2000. Blindsight: A conscious route to unconscious vision. *Current Biology* 10:R64-R67.

Gazzaniga, M.S., Fendrich, R. & Wessinger, C.M. 1994. Blindsight reconsidered. *Current Directions in Psychological Science* 3:93-96.

Graves, R.E. & Jones, B.S. 1992. Conscious visual perceptual awareness vs. non-conscious visual spatial localisation examined with normal subjects using possible analogues of blindsight and neglect. *Cognitive Neuropsychology* 9:487-508.

Guzeldere, G., Flanagan, O., & Hardcastle, V. 2000. The nature and function of consciousness: Lessons from blindsight. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.

Heywood, C.A., Cowey, A. & Newcombe, F. 1991. Chromatic discrimination in a cortically colour-blind observer. *European Journal of Neuroscience* 3:802-12.

Jackson, S. 2000. Perception, awareness and action: Insights from blindsight. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Klein, S.A. 1998. Double-judgment psychophysics for research on consciousness: Application to blindsight. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Kolb, F.C. & Braun, J. 1995. Blindsight in normal observers. *Nature* 377:336-8.

Marcel, A.J. 1998. Blindsight and shape perception: Deficit of visual consciousness or of visual function? *Brain* 121:1565-88.

Marshall, J.C. & Halligan, P.W. 1988. Blindsight and insight in visuospatial neglect. *Nature* 336:766-67.

Morgan, M.J., Mason, A.J.S. & Solomon, J.A. 1997. Blindsight in normal subjects? *Nature* 385:401-2.

Natsoulas, T. 1982. Conscious perception and the paradox of "blind-sight". In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Natsoulas, T. 1997. Blindsight and consciousness. *American Journal of Psychology* 110:1-33.

Paillard, J., Michel, F. & Stelmach, C.E. 1983. Localization without content: A tactile analogue of "blind sight". *Archives of Neurology* 40:548-51.

Place, U. 2000. Consciousness and the zombie within: A functional analysis of the blindsight evidence. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

- Sahraie, A., Weiskrantz, L., Barbur, J.L., Simmons, A., & Brammer, M. 1997. Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Sciences USA* 94:9406-9411.
- Stoerig, P. & Cowey, A. 1989. Wavelength sensitivity in blindsight. *Nature* 342:916-18.
- Stoerig, P. & Cowey, A. 1989. Wavelength sensitivity in blindsight. Wavelength sensitivity in blindsight. *Brain* 115:425-44.
- Stoerig, P. & Cowey, A. 1991. Increment threshold spectral sensitivity in blindsight: Evidence for colour opponency.
- Stoerig, P. & Cowey, A. 1993. Blindsight and perceptual consciousness: Neuropsychological aspects of striate cortical function. In (B. Gulyas, D. Ottoson, & P. Roland, eds) *Functional Organization of the Human Visual Cortex*. Pergamon Press.
- Stoerig, P. & Cowey, A. 1997. Blindsight in man and monkey. *Brain* 120:535-59.
- Stoerig, P. 1998. Varieties of vision: From blind responses to conscious recognition. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Tapp, J.D. 1997. Blindsight in hindsight. *Consciousness and Cognition* 6:67-74.
- Torjussen, T. 1978. Visual processing in cortically blind hemifields. *Neuropsychologia* 16:5-21.
- Vision, G. 1998. Blindsight and philosophy. *Philosophical Psychology* 11:137-59.
- Weiskrantz, L. 1986. *Blindsight: A Case Study and Implications*. Oxford University Press.
- Weiskrantz, L. 1995. Blindsight: Not an island unto itself. *Current Directions in Psychological Science* 4:146-151.
- Weiskrantz, L. 1995. Blindsight: Conscious vs. unconscious aspects. In (J. King & K. Pribram, eds) *Scale in Conscious Experience*. Lawrence Erlbaum.
- Weiskrantz, L. 1996. Blindsight revisited. *Current Opinion in Neurobiology* 6:215-220.
- Weiskrantz, L. 1997. *Consciousness Lost and Found*. Oxford University Press.
- Weiskrantz, L. 1998. Consciousness and commentaries. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Weiskrantz, L. & Cowey, A. 1970. Filling in the scotoma: A study of residual vision after striate cortex lesions in monkeys. (in (E. Stellar & J. Sprague, eds) *Progress in Physiological Psychology*. Academic Press.
- Weiskrantz, L., Barbur, J.L. & Sahraie, A. 1995. Parameters affecting conscious versus unconscious visual discrimination without V1. *Proceedings of the National Academy of Sciences USA* 92:6122-26.
- Weiskrantz, L. 2000. Blindsight: Implications for the conscious experience of emotion. In (R. Lane, L. Nadel, & G. Ahern, eds) *Cognitive Neuroscience of Emotion*. Oxford University Press.

Wessinger, C.M., Fendrich, R., Ptito, A., & Villemure, J. 1996. Residual vision with awareness in the field contralateral to a partial or complete functional hemispherectomy. *Neuropsychologia* 34:1129-1137.

Wessinger, C.M., Fendrich, R., Gazzaniga, M.S. 1997. Islands of residual vision in hemianopic patients. *Journal of Cognitive Neuroscience* 9:203-21.

Zeki S. 1996. The motion vision of the blind and the modularity of consciousness. *Transactions of the Medical Society of London* 112:11-18.

Zeki, S. & ffytche, D.H. 1998. The Riddoch syndrome: Insights into the neurobiology of conscious vision. *Brain* 121:25-45.

Zihl, J. 1980. "Blindsight": Improvement of visually guided eye movements by systematic practice in patients with cerebral blindness. *Neuropsychologia* 18:71-77.

Zihl, J. & von Cramon, D. 1980. Registration of light stimuli in the cortically blind hemifield and its effect on localization. *Behavior and Brain Research* 1:287-298.

Zihl, J. & Werth, R. 1984. Contributions to the study of "blindsight", parts I & II. *Neuropsychologia* 22:1-22.

6.1d Other Neuropsychological Disorders

Barr, W.B. 1998. Neurobehavioral disorders of awareness and their relevance to schizophrenia. In (X. Amador & A. David, eds) *_Insight and Psychosis_*. Oxford University Press

Bates, D. & Cartlidge, N. 1994. Disorders of consciousness. In (E. Critchley, ed) *_The Neurological Boundaries of Reality_*. Farrand.

Berti, A. & Rizzolatti, G. 1992. Visual processing without awareness: Evidence from unilateral neglect. *Journal of Cognitive Neuroscience* 4:345-51.

Bisiach, E., Luzzatti, C. & Perani, D. 1979. Unilateral neglect, representational schema, and consciousness. *Brain* 102:609-18.

Bisiach, E. & Rusconi, M.L. 1990. Breakdown of perceptual awareness in unilateral neglect. *Cortex* 26:643-49.

Bisiach, E. & Geminiani, G. 1991. Anosognosia related to hemiplegia and hemianopia. In (G. Prigatano & D. Schacter, eds) *_Awareness of Deficits after Brain Injury_*. Oxford University Press.

Bisiach, E. 1992. Understanding consciousness: Clues from unilateral neglect and related disorders. In (A. Milner & M. Rugg, eds) *_The Neuropsychology of Consciousness_*. Academic Press.

Bisiach, E. 1993. Mental representation in unilateral neglect and related disorders. *Quarterly Journal of Experimental Psychology* 46A:435-61.

Bisiach, E. & Berti, A. 1995. Consciousness in dyschiria. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.

Brady, J.P. & Lind, D.L. 1961. Experimental analysis of hysterical blindness. *Archives of General Psychiatry* 4:331-39.

De Giorgio, C.M. & Lew, M.F. 1991. Consciousness, coma, and the vegetative state: Physical basis and definitional character. *Issues in Law and Medicine* 6:361-371.

de Haan, E.H.F., Young, A.W., & Newcombe, F. 1987. Face recognition without awareness. *Cognitive Neuropsychology* 4:385-415.

de Renzi, E. 1986. Current issues in prosopagnosia. In (H. Ellis, M. Jeeves, F. Newcombe, & A. Young, eds) *_Aspects of Face Processing_*. Martinus Nijhoff.

Driver, J. & Vuilleumier, P. 2001. Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition* 79:39-88.

Engelien, A., Huber, W., Silbersweig, D., Frith, C., & Frachowiak, R. 2000. The neural correlates of 'deaf-hearing' in man. *Brain* 123:532-545.

Farah, M.J. 1990. *_Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision_*. MIT Press.

Farah, M.J. 1994. Perception and awareness after brain damage. *Current Opinion in Neurobiology* 4:252-55.

Farah, M.J. 1994. Visual perception and visual awareness after brain damage: A tutorial overview. In (C. Umiltà and M. Moscovitch, eds) *_Consciousness and Unconscious Information Processing: Attention and Performance 15_*. MIT Press.

Farah, M.J., O'Reilly, R.C. & Vecera, S.P. 1997. The neural correlates of perceptual awareness: Evidence from covert recognition in prosopagnosia. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Farah, M.J. & Feinberg, T.E. 1997. Consciousness of perception after brain damage. *Seminars in Neurology* 17:145-52.

Farah, M. & Feinberg, T. 2000. Disorders of perception and awareness. In (M. Farah & T. Feinberg, eds) *_Patient-based Approaches to Cognitive Neuroscience_*. MIT Press.

Farah, M. 2001. Consciousness. In (B. Rapp, ed) *_The Handbook of Cognitive Neuropsychology: What Deficits Reveal about the Human Mind_*. Psychology Press/Taylor & Francis.

Feinberg, T.E. 1997. Some interesting perturbations of the self in neurology. *Seminars in Neurology* 17:129-35.

Fredericks, J.A.M. 1969. Consciousness. In (P. Vinken & G. Bruyn, eds) *_Handbook of Clinical Neurology_*. North Holland.

Frith, C., Blakemore, S.J., & Wolpert, D. 2000. Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews* 31:357-363.

Galin, D. 1992. Theoretical reflections on awareness, monitoring, and self in relation on anosognosia. *Consciousness and Cognition* 1:152-62.

Giacino J.T. 1997. Disorders of consciousness: differential diagnosis and neuropathologic features. *Seminars in Neurology* 17:105-11.

Gibson, K.R. 1992. Toward an empirical basis for understanding consciousness and self-awareness. *Consciousness and Cognition* 1:163-68.

Grosz, H.J. & Zimmerman, J.A. 1965. Experimental analysis of hysterical blindness: A follow-up report and new experimental data. *Archives of General Psychiatry* 13:255-60.

Hellman, K.M. 1991. Anosognosia: Possible neuropsychological mechanisms. In (G. Prigatono & D. Schacter, eds) *_Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues_*. Oxford University Press.

Humphreys, G.W., Troscianko, T., Riddoch, M.J., & Boucart, M. 1992. Covert processing in different visual recognition systems. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Jehkonen, M., Ahonen, J., Dastidar, P., & Vilkkki, J. 2000. Unawareness of deficits after right hemisphere stroke: Double-dissociations of anosognosias. *Acta Neurologica Scandinavica* 102:378-384.

Jouvet, M. 1969. Coma and other disorders of consciousness. In (P. Vinken & G. Bruyn, eds) *Handbook of Clinical Neurology*. North Holland.

Katz, J. 2000. Individual differences in the consciousness of phantom limbs. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Kihlstrom, J.F. & Tobias, B.A. 1991. Anosognosia, consciousness, and the self. In (G. Prigatono & D. Schacter, eds) *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Oxford University Press.

Knight, R.T. & Grabowecky, M. 1995. Escape from linear time: Prefrontal cortex and conscious experience. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Koehler, S. & Moscovitch, M. 1997. Unconscious visual processing in neuropsychological syndromes: A survey of the literature and evaluation of models of consciousness. In (M. Rugg, ed) *Cognitive Neuroscience*. MIT Press.

Ladavas, E., Berti, A., & Farne, A. 2000. Dissociation between conscious and non-conscious processing in neglect. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Lane, R.D., Ahern, G.L., Schwartz, G.E. & Kaszniak, A.W. 1997. Is alexithymia the emotional equivalent of blindsight? *Biological Psychiatry* 42:834-44.

Light, G., & Braff, D. 2000. Do self-reports of perceptual anomalies reflect gating deficits in schizophrenia patients? *Biological Psychiatry* 47:463-467.

McGlynn, S.M. & Schacter, D.L. 1989. Unawareness of deficits in neuropsychological syndromes. *Journal of Clinical and Experimental Neuropsychology* 11:143-205.

Milner, A.D. & Rugg, M. (eds) 1991. *The Neuropsychology of Consciousness*. Academic Press.

Milner, A.D. 1991. Disorders of perceptual awareness: Commentary. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Newcombe, F. 1985. Neuropsychology of consciousness: A review of human clinical evidence. In (D. Oakley, ed) *Brain and Mind*. Methuen.

Porter R.J. 1991. Disorders of consciousness and associated complex behaviors. *Seminars in Neurology* 11:110-17.

Posner J.B. 1978. Coma and other states of consciousness: the differential diagnosis of brain death. *Annals of the New York Academy of Sciences* 315:215-27.

Prigatono, G.P. & Schacter, D.L. (eds) 1991. *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Oxford University Press.

Ramachandran, V.S. 1995. Anosognosia in parietal lobe syndrome. *Consciousness and Cognition* 4:22-51.

Rioch, D.M. 1954. Psychopathological and neuropathological aspects of consciousness. In (J. Delafresnaye, ed) *_Brain Mechanisms and Consciousness_*. Blackwell.

Schacter, D.L., McAndrews, M.P., and Moscovitch, M. 1986. Access to consciousness: Dissociations between implicit and explicit knowledge in neuropsychological syndromes. In (L. Weiskrantz, ed) *_Thought Without Language_*. Oxford University Press.

Schacter, D.L. 1990. Toward a cognitive neuropsychology of awareness: Implicit knowledge and anosognosia. *Journal of Clinical and Experimental Neuropsychology* 12:155-78.

Schiff, N. & Plum, F. 2000. The role of arousal and "gating" systems in the neurology of impaired consciousness. *Journal Of Clinical Neurophysiology* 17:438-452.

Teasdale G., Knill-Jones R., & van der Sande J. 1978. Observer variability in assessing impaired consciousness and coma. *Journal of Neurology, Neurosurgery and Psychiatry* 41:603-10.

Tononi, G. & Edelman, G. 2000. Schizophrenia and the mechanisms of conscious integration. *Brain Research Reviews* 31:391-400.

Tranel, D, & Damasio, A.R. 1988. Nonconscious face recognition in patients with prosopagnosia. *Behavioral Brain Research* 30:235-49.

Vaina, L.M. 1995. Akinetopsia, achromatopsia and blindsight: Recent studies on perception without awareness. *Synthese* 105:253-271.

van de Kelft E., Segnarbieux F., Candon E., Couchet P., Frerebeau P., Daures J.P. 1994. Clinical recovery of consciousness after traumatic coma. *Critical Care Medicine* 22:1108-13.

von Cramon, D. 1978. Consciousness and disturbances of consciousness. *Journal of Neurology* 219:1-13.

Vecera, S.P. & Gilds, K.S. 1997. What is it like to be a patient with apperceptive agnosia? *Consciousness and Cognition* 6:237-66.

Watson, R.T., Valenstein, E., Day, A., & Heilman, K.M. 1994. Posterior neocortical systems subserving awareness and neglect: Neglect associated with superior temporal sulcus but not area 7 lesions. *Archives of Neurology* 51:1014-1021

Weiskrantz, L. 1987. Neuropsychology and the nature of consciousness. In (C. Blakemore & S. Greenfield, eds) *_Mindwaves_*. Blackwell.

Weiskrantz, L. 1988. Some contributions of neuropsychology of vision and memory to the problem of consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Weiskrantz, L. 1990. Outlooks for blindsight: Explicit methodologies for implicit processes. *Proceedings of the Royal Society of London B* 239:247-78.

Weiskrantz, L. 1994. Neuropsychology and the nature of consciousness. In (H. Gutfreund & G. Toulouse, eds) *_Biology and Computation: A Physicist's Choice_*. World Scientific.

Yamadori, A. 1997. Body awareness and its disorders. In (M. Ito, Y.

Miyashita, & E.T. Rolls, eds) *Cognition, Computation, and Consciousness*.
Oxford University Press.

Young, A.W. 1994. Covert recognition. In (M. Farah & G. Ratcliff, eds) *The Neuropsychology of High-Level Vision*. Lawrence Erlbaum.

Young, A.W. 1994. Conscious and unconscious recognition of familiar faces. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance 15*. MIT Press.

Young, A.W. 1995. Neuropsychology of awareness. In (A. Revonsuo & M. Kampinnen, eds) *Consciousness in Philosophy and Cognitive Neuroscience*. Lawrence Erlbaum.

Young, A.W. 1995. Face recognition and awareness after brain injury. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Young, A.W. 1996. Dissociable aspects of consciousness. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Zappulla R.A. 1997. Epilepsy and consciousness. *Seminars in Neurology* 17:113-19.

6.1e Cerebral Hemispheres and Consciousness

Albert M.L., Silverberg R., Reches A., & Berman M. 1976. Cerebral dominance for consciousness. *Archives of Neurology* 33:453-4.

Austin, G., Hayward, W., & Rouhe, S. 1974. A note on the problem of conscious man and cerebral disconnection by hemispherectomy. In (M. Kinsbourne & W. Smith, eds) *Hemispheric Disconnection and Cerebral Function*. Charles C. Thomas.

Battro, A. 2001. *Half a Brain is Enough: The Story of Nico*. Cambridge University Press.

Baynes, K. & Gazzaniga, M. 2000. Consciousness, introspection, and the split-brain: The two minds/one body problem. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.

Beaumont, J. 1981. Split brain studies and the duality of consciousness. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 2*. Academic Press.

Bogen, J.E. 1968. The other side of the brain: An appositional mind. *Bulletin of the Los Angeles Neurological Society* 34:135-62.

Bogen, J.E. 1977. Further discussion of split brains and hemispheric capabilities. *British Journal for the Philosophy of Science* 28:281-6.

Dewitt, L. 1975. Consciousness, mind, self: The implications of the split-brain studies. *British Journal for the Philosophy of Science* 27:41-47.

Dimond, S.J. 1978. Depletion of awareness and double-simultaneous stimulation in split-brain man. *Cortex* 14:604-607.

Gazzaniga, M. 1977. On dividing the self: Speculations from brain research. *Excerpta Medica: Neurology* 434:233-44.

Gazzaniga, M.S., LeDoux, J.E., Wilson, D.H. 1977. Language, praxis, and the right hemisphere: Clues to some mechanisms of consciousness. *Neurology* 27:1144-1147.

Gazzaniga, M.S. 1995. Consciousness and the cerebral hemispheres. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Gazzaniga, M. & Miller, M. 2000. Testing Tulving: The split brain approach. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Harrington, A. 1985. Nineteenth-century ideas on hemisphere differences and "duality of mind." *Behavioral and Brain Sciences* 8:617-660.

Joseph, R. 1988. The right cerebral hemisphere: Emotion, music, visual-spatial skills, body-image, dreams, and awareness. *Journal of Clinical Psychology* 44:630-673.

Kavcic, V., Fei, R., Hu, S., & Doty, R. 2000. Hemispheric interaction, metacontrol, and mnemonic processing in split-brain macaques. *Behavioural Brain Research* 111:71-82.

Kurian, G., Santhakumari, K. 1990. Consciousness and the left cerebral hemisphere. *Journal of Indian Psychology* 8:33-36.

LeDoux, J.E., Wilson, D.H. & Gazzaniga, M.S. 1977. A divided mind: Observations of the conscious properties of the separated hemispheres. *Annals of Neurology* 2:417-21.

LeDoux, J.E., Wilson, D.H., & Gazzaniga, M.S. 1979. Beyond commissurotomy: Clues to consciousness. In (M. Gazzaniga, ed) *Handbook of Behavioral Neurobiology*, volume 2. Plenum Press.

LeDoux, J.E. 1986. Brain, mind, and language. In (D. Oakley, ed) *Brain and Mind*. Methuen.

Landis, T., Graves, R., & Goodglass, H. 1981. Dissociated awareness of manual performance on two different visual associative tasks: A "split-brain" phenomenon in normal subjects? *Cortex* 17:435-440.

Lishman, W.A. 1971. Emotion, consciousness, and will after brain bisection in man. *Cortex* 7:181-92.

Mackay, D.M. 1987. Divided brains -- divided minds? In (C. Blakemore and S. Greenfield, eds) *Mindwaves*. Blackwell.

Mark, V. 1996. Conflicting communication in a split-brain patient: Support for dual consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Marks, C. 1980. *Commissurotomy, Consciousness, and Unity of Mind*. MIT Press.

Miller, L. 1986. Some comments on cerebral hemispheric models of consciousness. *Psychoanalytic Review* 73:129-44.

Natsoulas, T. 1987. Consciousness and commissurotomy: 1. Spheres and Streams of consciousness. *Journal of Mind and Behavior* 8:435-468.

Natsoulas, T. 1988. Consciousness and commissurotomy: 2. Some Pertinencies for Intact Functioning. *Journal of Mind and Behavior* 9:515-548.

Natsoulas, T. 1991. Consciousness and commissurotomy: 3. Toward the improvement of alternative conceptions. *Journal of Mind and Behavior* 12:1-32.

Natsoulas, T. 1992. Consciousness and commissurotomy: 4. Three hypothesized dimensions of disconnected left-hemispheric consciousness. *Journal of Mind and Behavior* 13:37-67.

- Natsoulas, T. 1991. Consciousness and commissurotomy: 5. Concerning a hypothesis of normal dual consciousness. *Journal of Mind and Behavior* 14:179-202.
- Natsoulas, T. 1991. Consciousness and commissurotomy: 6. Evidence for normal dual consciousness. *Journal of Mind and Behavior* 16:181-205.
- Preilowski B. 1979. Self-recognition as a test of consciousness in left and right hemisphere of "split-brain" patients. *Activitas Nervosa Superior* 19 (supp):343-44.
- Puccetti, R. 1977. Bilateral organization of consciousness in man. *Annals of the New York Academy of Sciences* 299:448-58.
- Puccetti, R. 1981. The case for mental duality: Evidence from split-brain data and other considerations. *Behavioral and Brain Sciences* 4:93-123.
- Quen, J.M. (ed) 1986. *_Split Minds/Split Brains: Historical and Current Perspectives_*. New York University Press.
- Sergent, J. 1987. A new look at the human split brain. *Brain* 110:1375-92.
- Sperry, R.W. 1968. Hemisphere disconnection and unity in conscious awareness. *American Psychologist* 23:723-733.
- Sperry, R.W. 1977. Forebrain commissurotomy and conscious awareness. *Journal of Medicine and Philosophy* 2:101-26.
- Sperry, R.W., Zaidel, E., Zaidel, D. 1979. Self recognition and social awareness in the disconnected minor hemisphere. *Neuropsychologia* 17:153-166.
- Sperry, R.W. 1984. Consciousness, personal identity and the divided brain. *Neuropsychologia* 22:611-73.
- Trevarthen, C. 1974. Analysis of central activities that generate and regulate consciousness in commissurotomy patients. In (S. Dimond & J. Beaumont, eds) *_Hemisphere Function in the Human Brain_*. Elek.
- Wessinger, C.M., Fendrich, R., Ptito, A., & Villemure, J.G. 1996. Residual vision with awareness in the field contralateral to a partial or complete functional hemispherectomy. *Neuropsychologia* 34:1129-1137.
- Wilkes, K.V. 1978. Consciousness and commissurotomy. *Philosophy* 53:185-99.
- Zangwill, O.L. 1974. Consciousness and the cerebral hemispheres. In (S. Dimond & J. Beaumont, eds) *_Hemisphere Function in the Human Brain_*. Wiley.
- 6.1f Neural Timing and Consciousness (Libet, etc)

- Churchland, P.S. 1981. On the alleged backward referral of experience and its relevance to the mind-body problem. *Philosophy of Science* 48:165-81.
- Churchland, P.S. 1981. The timing of sensations: Reply to Libet. *Philosophy of Science* 48:492-7.
- Dennett, D.C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15:183-201.
- Elitzur, A. 1996. Time and consciousness: The uneasy bearing of relativity on the mind-body problem. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Glynn, I.M. 1990. Consciousness and time. *Nature* 348:477-79.

Green, C. & Gillett, G. 1995. Are mental events preceded by their physical causes? *Philosophical Psychology* 8:333-340.

Honderich, T. 1984. The time of a conscious sensory experience and mind-brain theories. *Journal of Theoretical Biology* 110:115-129.

Kiefer, M. & Spitzer, M. 2000. Time course of conscious and unconscious semantic brain activation. *Neuroreport* 11:2401-2407.

Libet, B. 1978. Neuronal vs. subjective timing for a conscious sensory experience. In (P. Buser & A. Rougeul-Buser, eds) *Cerebral Correlates of Conscious Experience*. Elsevier.

Libet, B. Wright, E.W, Feinstein, B. & Pearl, D.K. 1979. Subjective referral of the timing for a cognitive sensory experience. *Brain* 102:193-224.

Libet, B. 1981. The experimental evidence for subjective referral of a sensory experience backwards in time: Reply to P.S. Churchland. *Philosophy of Science*, 48, 182-97.

Libet, B. 1981. Timing of cerebral processes relative to concomitant conscious experiences in man. In (G. Adam, I. Meszaros & E.I. Banyai, eds), *Advances in Physiological Science*. Pergamon.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8:529-66.

Libet, B. 1985. Subjective antedating of a sensory experience and mind-brain theories: Reply to Honderich. *Journal of Theoretical Biology* 114:563-70.

Libet, B., Wright, E.W., Feinstein, B. & Pearl, D.K. 1992. Retroactive enhancement of a skin sensation by a delayed cortical stimulus in man: Evidence for delay of a conscious sensory experience. *Consciousness and Cognition* 1:367-75.

Libet, B. 1993. The neural time factor in conscious and unconscious events. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

Libet, B. 1993. *Neurophysiology of Consciousness: Selected Papers and New Essays*. Birkhauser.

Mele, A.R. 1997. Strength of motivation and being in control - Learning from Libet. *American Philosophical Quarterly* 34:319-32.

Rossi, E.L. 1988. Paradoxes of time, consciousness, and free will: Integrating Bohm, Jung, and Libet on ethics. *Psychological Perspectives* 19:50-55.

6.1g Neural Synchrony and Binding

Cotterill, R.M. & Nielsen, C. 1991. A model for cortical 40-hertz oscillations invokes inter-area interactions. *Neuroreport* 2:289-92.

Crick, F. & Koch, C. 1990. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2:263-275.

Damasio, A.R. 1989. The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation* 1:123-32.

Damasio, A.R. 1989. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recognition and recall. *Cognition*

3:25-62.

Damasio, A.R. 1990. Synchronous activation in multiple cortical regions: A mechanism for recall. *Seminars in the Neurosciences* 2:287-96.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., & Reitbock, H.J. 1988. Coherent oscillations: A mechanism for feature linking in the visual cortex. *Biological Cybernetics* 60:121-30.

Eckhorn, R., Reitbock, H.J., Arndt, M., & Dicke, P. 1989. A neural network for feature linking via synchronous activity: Results from cat visual cortex and from simulations. In (R. Cotterill, ed) *_Models of Brain Function_*. Cambridge University Press.

Engel, A.K., Konig, P. & Singer, W. 1991. Direct physiologic evidence for scene segmentation by temporal coding. *Proceedings of the National Academy of Sciences USA* 88:1936-40.

Engel, A.K., Konig, P. Kreiter, A.K. & Schillen, T.B. 1992. Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neurosciences* 15:218-26.

Engel, A.K., Fries, P., Konig, P., Brecht, M. & Singer, W. 1999. Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition* 8:128-51.

Engel, A.K., Fries, P., Konig, P., Brecht, M. & Singer, W. 1999. Does time help to understand consciousness? *Consciousness and Cognition* 8:260-68.

Fries, P., Roelfsema, P., Engel, A., & Singer, W. 1997. Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proceedings of the National Academy of Sciences USA* 94:12699-12704.

Gold, I. 1999. Does 40-Hz oscillation play a role in visual consciousness? *Consciousness and Cognition* 8:186-95.

Golledge, H.D.R., Hilgetag, C.C., & Tovee, M.J. 1996. Information processing: A solution to the binding problem. *Current Biology* 6:1092-95.

Gray, C.M., Konig, P., Engel, A.K. & Singer, W. 1992. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338:334-7.

Gray, C.M. 1994. Synchronous oscillations in neuronal systems: Mechanisms and functions. *Journal of Computational Neuroscience* 1:11-38.

Hardcastle, V.G. 1994. Psychology's "binding problem" and possible neurobiological solutions. *Journal of Consciousness Studies* 1:66-90.

Hardcastle, V.G. 1996. How we get there from here: Dissolution of the binding problem. *Journal of Mind and Behavior* 17:251-66.

Hardcastle, V.G. 1997. Consciousness and the neurobiology of perceptual binding. *Seminars in Neurology* 17:163-70.

Konig, P. & Engel, A.K. 1995. Correlated firing in sensory-motor systems. *Current Opinion in Neurobiology* 5:511-19.

Konig, P., Engel, A.K. & Singer, W. 1995. Relation between oscillatory activity and long-range synchronization in cat visual cortex. *Proceedings of the National Academy of Sciences USA* 92:290-94.

Konig, P., Engel, A.K., Roelfsema, P.R. & Singer, W. 1995. How precise is neural synchronization? *Neural Computation* 7:469-85.

- Llinas, R. & Ribary, U. 1998. Temporal conjunction in thalamocortical transactions. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.
- Nelson, J.I. 1995. Binding in the visual system. In (M. Arbib, ed) *_Handbook of Brain Theory and Neural Networks_*. MIT Press.
- Newman, J. & Grace, A.A. 1999. Binding across time: The selective gating of frontal and hippocampal systems modulating working memory and attentional states. *Consciousness and Cognition* 8:196-212.
- Prinzmetal, W. 1981. Principles of feature integration in visual perception. *Perception and Psychophysics* 30:330-40.
- Revonsuo, A. 1999. Binding and the phenomenal unity of consciousness. *Consciousness and Cognition* 8:173-85.
- Sauve, K. 1999. Gamma-band synchronous oscillations: Recent evidence regarding their functional significance. *Consciousness and Cognition* 8:213-24.
- Schillen, T.B. & Konig, P. 1994. Binding by temporal structure in multiple feature domains of an oscillatory neural network. *Biological Cybernetics* 5:397-405.
- Shastri, L. & Ajjanagadde, V. 1993. From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic binding using temporal synchrony. *Behavioral and Brain Sciences* 16:417-51.
- Sillito, A.M., Jones, H.E., Gerstein, G.L., & West, D.C. 1994. Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature* 369:479-82.
- Singer, W. 1993. Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology* 55:349-74.
- Singer, W. & Gray, C.M. 1995. Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* 18:555-86.
- Singer, W., Engel, A.K., Kreiter, A., Munk, M., & Roelfsema, P. 1997. Neuronal assemblies: Necessity, signature, and detectability. *Trends in Cognitive Sciences* 1:252-60.
- Steriade, M., McCormick, D.A., & Sejnowski, T.J. 1993. Thalamocortical oscillations in the sleeping and aroused brain. *Science* 262:679-85.
- Steriade, M. 1998. Corticothalamic networks, oscillations, and plasticity. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.
- Stryker, M. 1989. Is grandmother an oscillation? *Nature* 338:297-8.
- Treisman, A. 1980. A feature integration theory of attention. *Cognitive Psychology* 12:97-136.
- Treisman, A. 1996. The binding problem. *Current Opinion in Neurobiology* 6:171-8.
- von der Malsburg, C. 1995. Binding in models of perception and brain function. *Current Opinion in Neurobiology* 5:520-28.
- Usher, M. & Donnelly, N. 1998. Visual synchrony affects binding and segmentation in perception. *Nature* 394:179-82.

Wolfe, J.M. & Bennett, S.C. 1997. Preattentive object files: Shapeless bundles of basic features. *Vision Research* 37:25-43.

6.1h Consciousness and Anesthesia

Aitkenhead, A.R. 1993. Conscious awareness. In (P. Sebel, B. Bonke, & E. Winograd, eds) *_Memory and Awareness in Anesthesia_*. Prentice-Hall.

Alkire, M., Haier, R., & Fallon, J. 2000. Toward a unified theory of narcosis: Brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness. *Consciousness & Cognition* 9:370-386.

Andrade, J. 1995. Learning during anesthesia: A review. *British Journal of Psychology* 86:479-506.

Andrade, J. 1997. Investigations of hypesthesia: Using anesthetics to explore relationships between consciousness, learning, and memory. *Consciousness and Cognition* 5:562-80.

Andrade, J. & Jones, J.G. 1997. Awareness in anesthesia. In (G. Hall & M. Morgan, eds) *_Short Practice of Anesthesia_*. Chapman and Hall.

Andrade, J. 2000. NMDA receptor--mediated consciousness: A theoretical framework for understanding the effects of anesthesia on cognition? In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Bonke, B., Fitch, W. & Millar, K. (eds) 1990. *_Memory and Awareness In Anesthesia_*. Swets & Zeitlinger.

Bonke, B., Bovill, J.G., & Moerman, N. (eds) 1996. *_Memory and Awareness in Anesthesia III_*. Van Gorcum.

Caseley-Rondi, G., Merikle, P.M. & Bowers, K.S. 1994. Unconscious cognition in the context of general anesthesia. *Consciousness and Cognition* 3:166-95.

Cogliolo, P., Romano, V., Villani, R., & Galano, M. 1993. Effectiveness of Evans' technique for the evaluation of awareness. In (P. Sebel, B. Bonke, & E. Winograd, eds) *_Memory and Awareness in Anesthesia 2_*. Prentice-Hall.

Corner, M. 1976. The nature of consciousness: some persistent conceptual difficulties and a practical suggestion. *Progress in Brain Research* 45:471-5.

Eich, E., Reeves, J.L., & Katz, R.L. 1985. Anesthesia, amnesia, and the memory/awareness distinction. *Anesthesia and Analgesia* 64:1143-48.

Evans, J.M. 1987. Patient's experiences of awareness during general anesthesia. In (M. Rosen & J. Lunn, eds) *_Consciousness, Awareness, and Pain in General Anesthesia_*. Butterworths.

Flohr, H. 1995. An information-processing theory of anesthesia. *Neuropsychologia* 33:1169-80.

Flohr, H. 1998. On the mechanism of action of anesthetic agents. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Flohr, H. 2000. NMDA-receptor-mediated computational processes and phenomenal consciousness. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Franks, N.P. & Lieb, W.R. 1998. The molecular basis of general anesthesia:

Current ideas. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Franks, N. & Lieb, W. 2000. The role of NMDA receptors in consciousness: What we learn from anesthetic mechanisms? In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Ghoneim, M.M. & Block, R.I. 1992. Learning and consciousness during general anesthesia. *Anesthesiology* 76:279-305.

Hagan, S., Jibu, M. & Yasue, K. 1994. Consciousness and anesthesia: A hypothesis involving biophoton emission in the microtubular cytoskeleton of the brain. In (K. Pribram, ed) *Origins: Brain and Self-organization*. Lawrence Erlbaum.

Hameroff, S. 2001. Anesthesia: The "other side" of consciousness. *Consciousness and Cognition* 10:217-229.

Hill, D.S., & Hill, D.S. 1910. The loss and recovery of consciousness under anesthesia. *Psychological Bulletin* 7:77-83.

Jansen, C.K., Bonke, B., Klein, J. & Bezstarosti, J. 1990. Unconscious perception during balanced anesthesia? In (B. Bonke, W. Fitch, & K. Millar, 1990) *Memory and Awareness in Anesthesia*. Swets & Zeitlinger.

Jones, J.G. 1988. Awareness during anesthesia. *Anaesthesia Rounds*.

Kihlstrom, J.F. & Schacter, D.L. 1990. Anesthesia, amnesia, and the cognitive unconscious. In (B. Bonke, W. Fitch, & K. Millar, 1990) *Memory and Awareness in Anesthesia*. Swets & Zeitlinger.

Kihlstrom, J.F. & Couture, L.J. 1992. Awareness and information processing during general anesthesia. *Journal of Psychopharmacology* 6:410-17.

Kiviniemi K. 1994. Conscious awareness and memory during general anesthesia. *Aana Journal* 62:441-9.

Kulli, J. & Koch, C. 1991. Does anaesthesia cause loss of consciousness? *Trends in Neuroscience*, 14, 6-10.

Levinson, B.W. 1965. States of awareness during general anaesthesia. *British Journal of Anaesthesia* 37:544-546.

Lewis, S.A., Jenkinson, J. & Wilson, J. 1973. An EEG investigation of awareness during anaesthesia. *British Journal of Psychology* 64:413-5.

Merikle, P. M., & Daneman, M. 1996. Memory for unconsciously perceived events: Evidence from anesthetized patients. *Consciousness and Cognition* 5:525-541.

Merikle, P.M. & Daneman, M. 1996. Memory for events during anesthesia: A meta-analysis. In (B. Bonke, J.G. Bovill, & N. Moerman, eds) *Memory and Awareness in Anesthesia III*. Van Gorcum.

Moerman, N, Bonke, B. & Oosting, J. 1993. Awareness and recall during general anesthesia: Facts and feelings. *Anesthesiology* 79:454-64.

Mostert, J.W. 1975. States of awareness during general anesthesia. *Perspectives in Biology and Medicine* 19:68-76.

Munglani, R. & Jones, J.G. 1992. Sleep and general anesthesia as altered states of consciousness. *Journal of Psychopharmacology* 6:399-409.

Nikolinakos, D. 1994. General anesthesia, consciousness, and the skeptical challenge. *Journal of Philosophy* 2:88-104.

Plourde, G. 2001. Identifying the neural correlates of consciousness: Strategies with general anesthetics. *Consciousness and Cognition* 10:241-44.

Rosen, M., & Lunn, J.N. (eds) 1987. *_Consciousness, Awareness, and Pain in General Anesthesia_*. Butterworths.

Sebel, P.S., Bonke, B. & Winograd, E. (eds) 1993. *_Memory and Awareness in Anesthesia 2_*. Prentice-Hall.

Tinnin, L. 1994. Conscious forgetting and subconscious remembering of pain. *Journal of Clinical Ethics* 5:151-52.

Tracy, J. 1993. Awareness in the operating room: A patient's view. In (P. Sebel, B. Bonke, & E. Winograd, eds) *_Memory and Awareness in Anesthesia_*. Prentice-Hall.

Utting, J.E. 1987. Awareness: Clinical aspects. In (M. Rosen & J. Lunn, eds) *_Consciousness, Awareness, and Pain in General Anesthesia_*. Butterworths.

White, D.C. 1987. Anesthesia: A privation of the senses: An historical introduction and some definitions. In (M. Rosen & J. Lunn, eds) *_Consciousness, Awareness, and Pain in General Anesthesia_*. Butterworths.

Vickers, M.D. 1987. Detecting consciousness by clinical means. In (M. Rosen & J. Lunn, eds) *_Consciousness, Awareness, and Pain in General Anesthesia_*. Butterworths.

6.1i Foundational Issues

Baars, B.J. 2001. The brain basis of a "consciousness monitor":Scientific and medical significance. *Consciousness and Cognition* 10:159-164.

Bisiach, E. 1988. The (haunted) brain and consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Block, N. 1998. How to find the neural correlate of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness 1996_*. MIT Press.

Buck, R. 1993. What is this thing called subjective experience? Reflections on the neuropsychology of qualia. *Neuropsychology* 7:490-99.

Chalmers, D.J. 1998. On the search for the neural correlate of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Chalmers, D.J. 2000. What is a neural correlate of consciousness? In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Churchland, P.S. 1988. Reduction and the neurobiological basis of consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Churchland, P.S. 1994. Can neurobiology teach us anything about consciousness? *Proceedings and Addresses of the American Philosophical Association* 67:23-40.

Cleeremans, A. & Haynes, J. 1999. Correlating consciousness: A view from empirical science. *Revue Internationale de Philosophie* 3:387-420.

Cobb, S. 1952. On the nature and locus of mind. *Archives of Neurology and Psychiatry* 67:172-7.

- Creutzfeld, O.D. 1987. Inevitable deadlocks of the brain-mind discussion. In (B. Gulyas, ed) *The Brain-Mind Problem: Philosophical and Neurophysiological Approaches*. Leuven University Press.
- Farber, I.B. & Churchland, P.S. 1995. Consciousness and the neurosciences: Philosophical and theoretical issues. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Freeman, W.J. 1997. Three centuries of category errors in studies of the neural basis of consciousness and intentionality. *Neural Networks* 10:1175-83.
- Gillett, G. 1988. Consciousness and brain function. *Philosophical Psychology* 1:325-39.
- Gillett, G. 1995. Consciousness, thought, and neurological integrity. *Journal of Mind and Behavior* 16:215-33.
- Gloor, P. 1986. Consciousness as a neurological concept in epileptology: a critical review. *Epilepsia* 27:S14-26.
- Gordon, G., Maxwell, G. & Savodnik, I. (eds) 1976. *Consciousness and the Brain: A Scientific and Philosophical Inquiry*. Plenum.
- Gray, J. 1998. Creeping up on the hard question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Hamanaka, T. 1997. The concept of consciousness in the history of neuropsychiatry. *History of Psychiatry* 8:361-373.
- Hardcastle, V.G. 1995. *Locating Consciousness*. John Benjamins.
- Hardcastle, V.G. 1996. Discovering the moment of consciousness? I: Bridging techniques at work, & II: An ERP analysis of priming using novel visual stimuli. *Philosophical Psychology* 9:149-96.
- Hardcastle, V. 2000. How to understand the N in NCC. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Ivanitsky A.M. 1993. Consciousness: criteria and possible mechanisms. *International Journal of Psychophysiology* 14:179-87.
- Kinsbourne, M. 1997. What qualifies a representation for a role in consciousness? In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Lahav, R. 1993. What neuropsychology tells us about consciousness. *Philosophy of Science* 60:67-85.
- Lahav, R. 1997. The conscious and the nonconscious: Philosophical implications of neuropsychology. In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.
- Newton, N. 1991. Consciousness, qualia, and re-entrant signaling. *Behavior and Philosophy* 19:21.
- Niedermeyer E. 1994. Consciousness: Function and definition. *Clinical Electroencephalography* 25:86-93.
- Revonsuo, A. 1998. How to take consciousness seriously in cognitive neuroscience. *Communication and Cognition* 30:185-205.
- Revonsuo. A. 2000. Prospects for a scientific research program on

consciousness. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Schiller, F. 1952. Consciousness reconsidered. *Archives of Neurology and Psychiatry* 67:199-227.

6.1j Consciousness and Neuroscience, Misc -----

Arhem, P. 1996. Vertical information flow in the brain: on neuronal micro events and consciousness. *Biosystems* 38:191-98.

Bakhman, T. 2000. *_Microgenetic Approach to the Conscious Mind_*. John Benjamins.

Beck, H. 1976. Neuropsychological servosystems, consciousness, and the problem of embodiment. *Behavioral Science* 21:139-60.

Bernhaut, M., Gellhorn, E. & Rasmussen, A.T. 1953. Experimental contributions to the problem of consciousness. *Journal of Neurophysiology* 16:21-35.

Boitano, J. 1996. Edelmans's biological theory of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Borrett, D., Kelly, S., & Kwan, H. 2000. Phenomenology, dynamical neural networks and brain function. *Philosophical Psychology* 13:213-228.

Calvin, W. 1990. *_The Cerebral Symphony: Seashore Reflections on the Structure of Consciousness_*. Bantam.

Cowey, A. 1997. Current awareness: Spotlight on consciousness. *Developmental Medicine and Child Neurology* 39:54-62.

Creutzfeld, O.D. 1979. Neurophysiological mechanisms and consciousness. In *_Brain and Mind_* (Ciba Foundation Symposium 69). Elsevier.

Crick, F. & Koch, C. 1992. The problem of consciousness. *Scientific American* 267(3):152-60.

Crick, F. 1994. *_The Astonishing Hypothesis: The Scientific Search for the Soul_*. Scribners.

Damasio, A. 1999. *_The Feeling of What Happens: Body and Emotion in the Making of Consciousness_*. Harcourt Brace.

Dehaene, S. & Naccache, L. 2001. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79:1-37.

Delacour, J. 1995. An introduction to the biology of consciousness. *Neuropsychologia* 33:1061-1074.

Delacour, J. 1997. Neurobiology of consciousness: An overview. *Behavioural Brain Research* 85:127-141.

Delafresnaye, J.F. (ed) 1954. *_Brain Mechanisms and Consciousness_*. Blackwell.

Desmedt J.E. & Tomberg, C. 1995. Consciousness. *Electroencephalography and Clinical Neurophysiology, Supplement* 44:227-34.

Donald, M. 1995. The neurobiology of human consciousness: An evolutionary approach. *Neuropsychologia* 33:1087-1102.

Donchin, E., McCarthy, G., Kutas, M. & Ritter, W. 1983. Event-related brain potentials in the study of consciousness. In (R. Davidson, S. Schwartz, & D. Shapiro, eds) *_Consciousness and Self-Regulation_*. Plenum Press.

Donnelly G.F. 1982. Consciousness: the brain and self-regulation modalities. *Topics in Clinical Nursing* 3:13-20.

Doty, R.W. 1975. Consciousness from neurons. *Acta Neurobiologiae Experimentalis* 35:791-804.

Eccles, J.C. (ed) 1966. *_Brain and Conscious Experience_*. Springer.

Eccles, J.C. 1974. Cerebral activity and consciousness. In (F. Ayala & T. Dobzhansky, eds) *_Studies in the Philosophy of Biology_*. University of California Press.

Eccles, J.C. 1987. The effect of silent thinking on the cerebral cortex. In (B. Gulyas, ed) *_The Brain-Mind Problem: Philosophical and Neurophysiological Approaches_*. Leuven University Press.

Edelman, G.M. 1989. *_The Remembered Present: A Biological Theory of Consciousness_*. Basic Books.

Edelman, G.M. 1992. *_Bright Air, Brilliant Fire: On the Matter of the Mind_*. Penguin.

Fessard, A.E. 1952. Mechanisms of nervous integration and conscious experience. In (J. Delafresnaye, ed) *_Brain Mechanisms and Consciousness_*. Blackwell.

Frith, C.D. 1992. Consciousness, information processing, and the brain. *Journal of Psychopharmacology* 6:436-40.

Frith, C.D. 1996. The role of the prefrontal cortex in self-consciousness: the case of auditory hallucinations. *Philosophical Transactions of the Royal Society of London* B351:1505-12.

Gastaut, H. 1954. The brain stem and cerebral electrogenesis in relation to consciousness. In (J. Delafresnaye, ed) *_Brain Mechanisms and Consciousness_*. Blackwell.

Gazzaniga, M. 1988. Brain modularity: Toward a philosophy of conscious experience. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Globus, G., Maxwell, G. & Savodnik, I. (eds) 1975. *_Consciousness and the Brain_*. Plenum Press.

Greenfield, S. 1995. *_Journey to the Centers of the Mind_*. W.H. Freeman.

Jasper, H. & Shagass, C. 1941. Conscious time judgments related to conditioned time intervals and voluntary control of the alpha rhythm. *Journal of Experimental Psychology* 28:503-508.

Kety, S.S. 1952. Consciousness and the metabolism of the brain. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the Third Conference_*. Josiah Macy Foundation.

Kinsbourne, M. 1995. Models of consciousness: Serial or parallel in the brain? In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.

Kokoszka, A. 1993. Information metabolism as a model of consciousness. *International Journal of Neuroscience* 68:165-77.

- Lin, S., Tsai, Y., & Liou, C. 1993. Conscious mental tasks and their EEG signals. *Medical and Biological Engineering and Computing* 31:421-26.
- Luria, A. 1976. The human brain and conscious activity. In Schwartz & Shapiro 1978.
- Mitterauer B. 1998. An interdisciplinary approach towards a theory of consciousness. *Biosystems* 45:99-121.
- Pare, D. & Llinas, R. 1995. Conscious and pre-conscious processes as seen from the standpoint of sleep-waking cycle neurophysiology. *Neuropsychologia* 33:1155-1168.
- Penfield, W. 1975. *The Mystery of the Mind*. Princeton University Press.
- Petty, P.G. 1998. Consciousness: A neurosurgical perspective. *Journal of Consciousness Studies* 5:86-96.
- Picton, T.W. & Stuss, D.T. 1994. Neurobiology of conscious experience. *Current Opinion in Neurobiology* 4:256-65.
- Pribram, K.H. 1990. Brain and consciousness: A wealth of data. In (E. John, ed), *Machinery of the Mind: Data, Theory, and Speculations about Higher Brain Function*. Boston: Birkhauser.
- Ramachandran, V.S. & Hirstein, W. 1998. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies* 4:429-57.
- Rose, S. 1973. *The Conscious Brain*. Paragon House.
- Rudell, A.P. & Hua, J. 1996. The recognition potential and conscious awareness. *Electroencephalography and Clinical Neurophysiology* 98:309-18.
- Salazar, A., Grafman J., Vance S., Dillon J.D., & Ludlow, C. 1986. Consciousness and amnesia after penetrating head injury: neurology and anatomy. *Neurology* 36:178-87.
- Singer, W. 2000. Phenomenal awareness and consciousness from a neurobiological perspective. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.
- Simonov, P.V. 1994. Consciousness and the brain. *Neuroscience and Behavioral Physiology* 24:234-38.
- Travis F.T. & Orme-Johnson D.W. 1989. Field model of consciousness: EEG coherence changes as indicators of field effects. *International Journal of Neuroscience* 49:203-11.
- Trevarthen, C. 1979. The tasks of consciousness: How could the brain do them? In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.
- Turner B.H. & Knapp M.E. 1995. Consciousness: A neurobiological approach. *Integrative Physiological and Behavioral Science* 30:151-6.
- Walter, W.G. 1954. Theoretical properties of diffuse projection systems in relation to behaviour and consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.
- Webb, A.C. 1970. Consciousness and the cerebral cortex. *British Journal of Anaesthesia* 55:209-19.
- Zeman, A.Z.J., Grayling, A.C. & Cowey, A. 1997. Contemporary theories of consciousness. *Journal of Neurology, Neurosurgery, and Psychiatry* 62:549-552.

6.2 Consciousness and Psychology

6.2a Cognitive Models of Consciousness

Allport, D.A. 1979. Conscious and unconscious cognition: A computational metaphor for the mechanism of attention and integration. In (L. Nilsson, ed) Perspectives on Memory Research.

Baars, B.J. 1983. Conscious contents provide the nervous system with coherent, global information. In (R. Davidson, G. Schwartz, & D. Shapiro, eds) Consciousness and Self-Regulation. Plenum.

Baars, B.J. 1988. A Cognitive Theory of Consciousness. Cambridge University Press.

Baars, B.J. 1993. How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity? In Experimental and Theoretical Studies of Consciousness (Ciba Foundation Symposium 174).

Baars, B.J., Fehling, M.R., LaPolla, M., & McGovern, K. 1997. Consciousness creates access: Conscious goal images recruit unconscious action routines, but goal competition serves to "liberate" such routines, causing predictable slips. In (J. Cohen & J. Schooler, eds) Scientific Approaches to Consciousness. Lawrence Erlbaum.

Baars, B.J. 1997. In the Theater of Consciousness: The Workspace of the Mind. Oxford University Press.

Baars, B.J. 1997. In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies* 4:292-309.

Baars, B.J. 1998. Metaphors of consciousness and attention in the brain. *Trends in Neurosciences* 21:58-62.

Baars, B.J. & McGovern, K. 1996. Cognitive views of consciousness: What are the facts? How can we explain them? In (M. Velmans, ed) The Science of Consciousness. Routledge.

Bechtel, W. 1995. Consciousness: Perspectives from symbolic and connectionist AI. *Neuropsychologia*.

Brown, R.A. 1997. Consciousness in a self-learning, memory-controlled, compound machine. *Neural Networks* 10:1333-85.

Browne, C., Evans, R., Sales, N., & Aleksander, I.L. 1997. Consciousness and neural cognizers: A review of some recent approaches. *Neural Networks* 10:1303-1316.

Burks, A.W. 1986. An architectural theory of functional consciousness. In (N. Rescher, ed) Current Issues in Teleology. University Press of America.

Cabanac, M. 1996. On the origin of consciousness, a postulate, and its corollary. *Neuroscience and Biobehavioral Reviews* 20:33-40.

Cam, P. 1989. Notes toward a faculty theory of cognitive consciousness. In (P. Slezak, ed) Computers, Brains and Minds. Kluwer.

Carr, T.H. 1979. Consciousness in models of human information processing: Primary memory, executive control, and input regulation. In (G. Underwood &

- R. Stevens, eds) *_Aspects of Consciousness, Volume 1_*. Academic Press.
- Cotterill, R.M.J. 1996. Prediction and internal feedback in conscious perception. *Journal of Consciousness Studies* 3:245-66.
- Cotterill, R.M.J. 1997. On the mechanism of consciousness. *Journal of Consciousness Studies* 4:231-48.
- Cotterill, R.M.J. 1997. Navigation, consciousness and the body/mind "problem". *Psyke and Logos*, 18:337-341.
- Farrell, B.A. 1970. The design of a conscious device. *Mind* 79:321-46.
- Gregory R.L. 1984. Is consciousness sensational inferences? *Perception* 13:641-6.
- Hardcastle, V.G. 1995. A critique of information processing theories of consciousness. *Minds and Machines* 5:89-107.
- Harnad, S. 1982. Consciousness: An afterthought. *Cognition and Brain Theory* 5:29-47.
- Harth, E. 1993. *_The Creative Loop: How the Brain Makes a Mind_*. Addison Wesley.
- Harth, E. 1995. The sketchpad model: A theory of consciousness, perception, and imagery. *Consciousness and Cognition* 4:346-68.
- Harth, E. 1996. Self-referent mechanisms as the neuronal basis of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.
- Jackendoff, R. 1987. *_Consciousness and the Computational Mind_*. MIT Press.
- John, E.R. 1976. A model of consciousness. In (G. Schwartz & D. Shapiro, eds) *_Consciousness and Self-Regulation_*. Plenum Press.
- Johnson, M.K. & Reeder, J.A. 1997. Consciousness as meta-processing. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.
- Johnson-Laird, P. 1983. A computational analysis of consciousness. *Cognition and Brain Theory* 6:499-508. Also in (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.
- Kawato, M. 1997. Bidirectional theory approach to consciousness. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *_Cognition, Computation, and Consciousness_*. Oxford University Press.
- Lauro-Grotto, R., Reich, S. & Virasoro, M.A. 1997. The computational role of conscious processing in a model of semantic memory. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *_Cognition, Computation, and Consciousness_*. Oxford University Press.
- Lloyd, D. 1995. Consciousness: A connectionist manifesto. *Minds and Machines* 5:161-85.
- Lloyd, D. 1996. Consciousness, connectionism, and cognitive neuroscience: A meeting of the minds. *Philosophical Psychology* 9:61-78.
- Mathis, D.W. & Moxer, M. 1995. On the computational utility of consciousness. In (G. Tesauro, D. Touretzky, & T. Leen, eds) *_Advances in Neural Information Processing Systems 7_*. MIT Press.

Michie, D. 1994. Consciousness as an engineering issue (Parts 1 and 2).
Journal of Consciousness Studies 1:192-95, 2:52-66.

Norretranders, T. 1991. *The User Illusion: Cutting Consciousness Down to Size*. Viking Penguin.

O'Brien, G. & Opie, J. 1998. A connectionist theory of phenomenal experience.
Behavioral and Brain Sciences 22:127-48.

Oatley, K. 1981. Representing ourselves: Mental schemata, computational metaphors, and the nature of consciousness. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 2*. Academic Press.

Parsons, T. 1953. Consciousness and symbolic processes. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Fourth Conference*. Josiah Macy Foundation.

Phaf, R.H. & Wolters, G. 1997. A constructivist and connectionist view on conscious and nonconscious processes. *Philosophical Psychology* 10:287-307.

Restian, A. 1981. Informational analysis of consciousness. *International Journal of Neuroscience* 13:229-37.

Revonsuo, A. 1993. Cognitive models of consciousness. In (M. Kamppinen, ed) *Consciousness, Cognitive Schemata, and Relativism*. Kluwer.

Roberts, H. 1968. Consciousness in animals and automata. *Psychological Reports* 22:1226-28.

Rolls, E.T. 1997. Consciousness in neural networks? *Neural Networks* 10:1227-1303.

Schacter, D.L. 1989. On the relation between memory and consciousness: Dissociable interactions and conscious experience. In (H. Roediger & F. Craik, eds) *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*.

Schneider, W. & Pimm-Smith, M. 1997. Consciousness as a message-aware control mechanism to modulate cognitive processing. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Shallice, T. 1972. Dual functions of consciousness. *Psychological Review* 79:383-93.

Shallice, T. 1978. The dominant action system: An information-processing approach to consciousness. In (K.S. Pope & J.L. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Shallice, T. 1988. Information-processing models of consciousness: possibilities and problems. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Sommerhoff G. & MacDorman K. 1994. An account of consciousness in physical and functional terms: A target for research in the neurosciences. *Integrative Physiological and Behavioral Science* 29:151-81.

Sommerhoff, G. 1996. Consciousness as an internal integrating system. *Journal of Consciousness Studies* 3:139-57.

Strehler, B.L. 1989. Monitors: key mechanisms and roles in the development and aging of the consciousness and self. *Mechanisms of Ageing and Development* 47:85-132.

Sun, R. 1997. Learning, action, and consciousness: A hybrid approach toward modeling consciousness. *Neural Networks* 10:1317-33.

Sviderskaya, N.E. 1991. Consciousness and information selection. *Neuroscience and Behavioral Physiology* 21:526-31.

Taylor, J.G. 1996. Modeling what it is like to be. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Taylor, J.G. 1996. A competition for consciousness? *Neurocomputing* 11:271-96.

Taylor, J.G. 1997. Neural networks for consciousness. *Neural Networks* 10:1207-27.

Taylor, J.G. & Mueller-Gaertner, H. 1997. Non-invasive analysis of awareness. *Neural Networks* 10:1185-1194.

Werbos, P. 1997. Optimization: A foundation for understanding consciousness. In (D. Levine & W. Elsberry, eds) *Optimality in Biological and Artificial Networks?*. Lawrence Erlbaum.

6.2b Unconscious Perception

Adams, J.K. 1957. Laboratory studies of behavior without awareness. *Psychological Bulletin* 54:383-405.

Balay, J. & Shevrin, H. 1988. The subliminal psychodynamic activation method: A critical review. *American Psychologist* 43:161-74.

Bar, M. 2000. Conscious and nonconscious processing of visual object identity. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Bargh, J.A. 1992. Does subliminality matter to social psychology? Awareness of the stimulus versus awareness of its influence. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.

Bhalla, M. & Proffitt, D. 2000. Geographical slant perception: Dissociation and coordination between explicit awareness and visually guided actions. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Bornstein, R.F. 1989. Exposure and affect: Overview and meta-analysis of research 1968-1987. *Psychological Bulletin* 106:265-89.

Bornstein, R.F. 1992. Subliminal mere exposure effects. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.

Bornstein, R.F. & Pittman, T.S. 1992. *Perception without Awareness: Cognitive, Clinical, and Social Perspectives*. Guilford.

Bowers, K.S. 1982. On being unconsciously influenced and informed. In (K. Bowers & D. Meichenbaum, eds) *The Unconscious Reconsidered*. Wiley.

Cheesman, J. & Merikle, P.M. 1984. Priming with and without awareness. *Perception and Psychophysics* 36:387-95.

Cheesman, J. & Merikle, P.M. 1986. Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology* 40:343-67.

Damian, M. 2001. Congruity effects evoked by subliminally presented primes: Automaticity rather than semantic processing. *Journal of Experimental Psychology: Human Perception & Performance* 27:154-165.

Debner, J.A. & Jacoby, L.L. 1994. Unconscious perception: Attention,

awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20:304-17.

Dixon, N.F. 1971. *_Subliminal Perception: The Nature of a Controversy_*. McGraw-Hill.

Dixon, N.F. & Henley, S.H.A. 1980. Without awareness. In (M. Jeeves, ed) *_Psychology Survey 3_*. Allen and Unwin.

Doyle, J.R. 1990. Detectionless processing with semantic activation? A footnote to Greenwald, Klinger, and Liu (1989). *Memory and Cognition* 18:428-9.

Eagle, M. 1959. The effects of subliminal stimuli of aggressive content upon conscious cognition. *Journal of Personality* 27:578-600.

Erdelyi, M.H. 1970. Recovery of unavailable perceptual input. *Cognitive Psychology* 1:99-113.

Eriksen, C.W. 1956. An experimental analysis of subception. *American Journal of Psychology* 69:625-34.

Eriksen, C.W. 1956. Subception: Fact or artifact? *Psychological Review* 63:74-80.

Eriksen, C.W. 1960. Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review* 67:279-300.

Farah, M.J., Monheit, M.A. & Wallace, M.A. 1991. Unconscious perception of "extinguished" visual stimuli: Reassessing the evidence. *Neuropsychologia* 29:949-58.

Fowler, C.A., Woldford, G., Slade, R. & Tassinary, L. 1981. Lexical access with and without awareness. *Journal of Experimental Psychology: General* 110:341-62.

Fuhrer, M.J. & Eriksen, C.W. 1960. The unconscious perception of the meaning of verbal stimuli. *Journal of Abnormal and Social Psychology* 61:432-9.

Goldiamond, I. 1958. Indicators of perception: 1. Subliminal perception, subception, unconscious perception: An analysis in terms of psychophysical indicator methodology. *Psychological Bulletin* 55:373-411.

Greenwald, A.G., Klinger, M.R., & Liu, T.J. 1989. Unconscious processing of dichoptically masked words. *Memory and Cognition* 17:35-47.

Greenwald, A.G., Spangenberg, E., Pratkanis, A.R., & Eskenazi, J. 1991. Double blind tests of subliminal self-help audiotapes. *Psychological Science* 2:119-22.

Greenwald, A.G., Klinger, M.R. & Schuh, E.S. 1995. Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General* 124:22-42.

Greenwald, A.G. & Draine, S. 1997. Do subliminal stimuli enter the mind unnoticed?: Tests with a new method. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Hardaway, R.A. 1990. Subliminally activated symbiotic fantasies: Facts and artifacts. *Psychological Bulletin* 107:177-95.

Henley, S.H. 1984. Unconscious perception revisited: A comment on Merikle (1992). *Bulletin of the Psychonomic Society* 22:121-4.

Holender, D. 1986. Semantic activation without conscious identification in

dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences* 9:1-23.

Kemp-Wheeler, S.M. & Hill, A.B. 1988. Semantic priming without awareness: Some methodological considerations and implications. *Quarterly Journal of Experimental Psychology* 40:671-92.

Khurana, B. 2000. Face representation without conscious processing. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Kihlstrom, J.F. 1996. Perception without awareness of what is perceived, learning without awareness of what is learned. In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.

Kihlstrom, J.F., Barnhardt, T.M. & Tataryn, D.J. 1992. Implicit perception. In (R. Bornstein & T. Pittman, eds) *_Perception without Awareness_*. Guilford.

Klauer, K. & Greenwald, A. 2000. Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception & Performance* 26:1506-1508.

Kostandov, E.A. 1994. Subsensory reactions and the problem of unconscious perception. *Sensory Systems* 7:149-53.

Krosnick, J.A., Betz, A.L., Jussim, L.J. & Lynn, A.R. 1992. Subliminal conditioning of attitudes. *Personality and Social Psychology Bulletin* 18:152-62.

Kunst-Wilson, W.R. & Zajonc, R.B. 1980. Affective discrimination of stimuli that cannot be recognized. *Science* 207:557-58.

Kunzendorf, R.G. 1985. Subconscious percepts as "unmonitored" percepts: An empirical study. *Imagination, Cognition and Personality* 4:365-73.

Lazarus, R.S. & McCleary R.A. 1983. Autonomic discrimination without awareness: A study of subception. *Psychological Review* 58:113-22.

Lewicki, P., Hill, T. & Czyewska, M. 1992. Nonconscious acquisition of information. *American Psychologist* 47:792-801.

Lewis, J.L. 1970. Semantic processing of unattended messages using dichotic listening.

MacLeod, C. 1998. Implicit perception: Perceptual processing without awareness. In (K. Kirsner, G. Speelman, eds) *_Implicit and Explicit Mental Processes_*. Lawrence Erlbaum.

Marcel, A.J. 1983. Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology* 15:197-237.

Marcel, A.J. 1983. Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology* 15:238-300.

Merikle, P.M. 1982. Unconscious perception revisited. *Perception and Psychophysics* 31:298-301.

Merikle, P.M. & Reingold, E.M. 1990. Recognition and lexical decision without detection: Unconscious perception? *Journal of Experimental Psychology: Human Perception and Performance* 16:574-83.

Merikle, P.M. 1992. Perception without awareness: Critical issues. *American Psychologist* 47:792-5.

- Merikle, P.M & Reingold, E.M. 1992. Measuring unconscious processes. In (R. Bornstein & T. Pittman, eds) *_Perception without Awareness_*. Guilford.
- Merikle, P.M., Joordens, S. & Stolz, J.A. 1995. Measuring the relative magnitude of unconscious influences. *Consciousness and Cognition* 4:422-39.
- Merikle, P.M. & Daneman, M. 1997. Psychological investigations of unconscious perception. *Journal of Consciousness Studies*.
- Merikle, P. & Daneman, M. 2000. Conscious vs. unconscious perception. In (M. Gazzaniga, ed) *_The New Cognitive Neurosciences: 2nd Edition_*. MIT Press.
- Merikle, P.M., D. Smilek, & J.D. Eastwood 2001. Perception without awareness: Perspectives from cognitive psychology. *Cognition* 79:115-34.
- Miller, J. 2000. Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception & Performance* 26:1461-1477.
- Moore, T.E. 1992. Subliminal perception: Facts and fallacies. *Skeptical Inquirer* 16:273-81.
- Neuberg, S.L. 1988. Behavioral implications of information presented outside of conscious awareness: The effect of subliminal presentation of trait information on behavior in the Prisoner's Dilemma game. *Social Cognition* 6:207-30.
- Nolan, K.A. & Caramazza, A. 1982. Unconscious perception of meaning: A failure to replicate. *Bulletin of the Psychonomic Society* 20:23-26.
- Peirce, C.S. & Jastrow, J. 1884. On small differences in sensation. *Memoirs of the National Academy of Sciences* 3:75-83.
- Pisella, L. & Rosetti, Y. 2000. Interaction between conscious identification and non-conscious sensory-motor processing: Temporal constraints. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.
- Poppel, E., Held, R. & Frost, D. 1973. Residual function after brain wounds involving the central visual pathways in man. *Nature* 243:295-96.
- Pratkanis, A.R. & Greenwald, A.G. 1988. Recent perspectives on unconscious processing: Still no marketing applications. *Psychology and Marketing* 5:337-53.
- Purcell, D.G., Stewart, A.L. & Stanovich, K.K. 1983. Another look at semantic priming without awareness. *Perception and Psychophysics* 34:65-71.
- Reingold, E.M. & Merikle, P.M. 1988. Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics* 44:563-575.
- Reingold, E. & Merikle, P. 1991. Theory and measurement in the study of unconscious processes. *Mind and Language* 5:9-28.
- Schmidt, T. 2000. Visual perception without awareness: priming responses by color. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.
- Shevrin, H. 1992. Unconscious perception, memory, and consciousness: Cognitive and dynamic perspectives. In (R. Bornstein & T. Pittman, eds) *_Perception without Awareness_*. Guilford.
- Silverman, L.H. & Weinberger, J. 1985. Mommy and I are one: Implications for psychotherapy. *American Psychologist* 40:1296-1308.

Theus, K.T. 1994. Subliminal advertising and the psychology of processing unconscious stimuli: A review of research. *Psychology and Marketing* 11:271-290.

Young, A. & Ellis, H. 2000. Overt and covert face recognition. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

Weinberger, J. & Hardaway, R. 1990. Separating science from myth in subliminal psychodynamic activation. *Clinical Psychological Review* 10:727-56.

6.2c Unconscious Processes

Abrams, R. & Greenwald, A. 2000. Parts outweigh the whole (word) in unconscious analysis of meaning. *Psychological Science* 11:118-124.

Bridgeman, B. 1992. Conscious vs unconscious processes: The case of vision. *Theory and Psychology* 2:73-88.

Buchner, A. 1997. Consciousness, intention, and the process dissociation procedure. *Sprache and Kognition* 16:176-182.

Cowan, N., Stadler, M.A. 1996. Estimating unconscious processes: Implications of a general class of models. *Journal of Experimental Psychology: General* 125:195-200.

Dixon, N.F. 1981. *_Preconscious Processing_*. Wiley.

Erdelyi, M.H. 1974. A new look at the New Look: Perceptual defense and vigilance. *Psychological Review* 81:1-25.

Erdelyi, M.H. 1992. Psychodynamics and the unconscious. *American Psychologist* 47:784-87.

Field, A. 2000. I like it, but I'm not sure why: Can evaluative conditioning occur without conscious awareness? *Consciousness & Cognition* 9:13-36.

Ford, T. & Thompson, E. 2000. Preconscious and postconscious processes underlying construct accessibility effects: An extended search model. *Personality & Social Psychology Review* 4:317-336.

Gaito, J. 1964. Stages of perception, unconscious processes, and information extraction. *Journal of General Psychology* 70:183-197.

Greenwald, A.G. 1992. New Look 3: Unconscious cognition reclaimed. *American Psychologist* 47:766-79.

Hilgard, E.R. 1958. *_Unconscious Processes and Man's Rationality_*. University of Illinois Press.

Hoffman, R. 1997. What neural network studies suggest regarding the boundary between conscious and unconscious mental processes. In (D. Stein, ed) *_Cognitive Science and the Unconscious_*. American Psychiatric Press

Hommel, B. 2000. Intentional control of automatic stimulus-response translation. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

Imanaka, K. & Abernethy, B. 2000. Distance-location interference in movement reproduction: An interaction between conscious and unconscious processing? In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

Jacoby, L.L., Toth, J.P., Lindsay, D.S., Debner, J.A. 1992. Lectures for a layperson: Methods for revealing unconscious processes. In (R. Bornstein & B. Pittman, eds) *Perception without Awareness: Cognitive, Clinical, and Social Perspectives*. Guilford Press.

Kihlstrom, J.F. 1984. Conscious, subconscious, unconscious: A cognitive perspective. In (K.S. Bowers & D. Meichenbaum, eds) *The Unconscious Reconsidered*. Wiley.

Kihlstrom, J.F. 1987. The cognitive unconscious. *Science* 237:1445-1452.

Kihlstrom, J.F. 1990. The psychological unconscious. In (L. Pervin, ed) *Handbook of Personality: Theory and Research*. Guilford Press.

Kihlstrom, J.F., Barnhardt, T.M. & Tattryn, D.J. 1992. The psychological unconscious: Found, lost, and regained. *American Psychologist* 47:788-91.

Kihlstrom, J.F. 1995. The rediscovery of the unconscious mind. In (H. Morowitz & J. Singer, eds) *The Mind, the Brain, and Complex Adaptive Systems*. Addison-Wesley.

Kihlstrom, J.F. 1996. Unconscious processes in social interaction. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Klinger, M., Burton, P., & Pitts, G. 2000. Mechanisms of unconscious priming: Response competition, not spreading activation. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 26:441-455.

Lewicki, P. 1986. *Nonconscious Social Information Processing*. Academic Press.

Lewicki, P. & Hill, T. 1987. Unconscious processes as explanations of behavior in cognitive, personality, and social psychology. *Personality and Social Psychology Bulletin* 13:355-362.

Loftus, E.F. & Klinger, M.R. 1992. Is the unconscious smart or dumb? *American Psychologist* 47:761-65.

Miller, J.G. 1951. Unconscious processes and perception. In (R. Blake & G. Ramsey, eds) *Perception*.

Miller, J.G. 1952. The experimental study of unconscious processes. In (M. Reymert, ed) *Feelings and Emotions*.

Perner, J. & Clements, W. 2000. From an implicit to an explicit "theory of mind". In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Peterfreund, E. & Schwartz, J.T. 1971. Information processing and the nature of conscious and unconscious processes. *Psychological Issues* 7:219-29.

Posner, M.I. 1991. Recent experimental studies of conscious and unconscious processes. In (M. Posner, B. Dwivedi, & I. Singh, eds) *Contemporary Approaches to Cognitive Psychology*. Rishi Publications.

Radil, T., Radilova, J., Bozkov, V., & Bohdanecky, Z. 1981. Unconscious and conscious processes during visual perception. *Acta Neurobiologiae Experimentalis* 41:565-572.

Revonsuo, A., Johanson, M., Wedlund, J., & Chaplin, J. 2000. The zombies among us: Consciousness and automatic behaviour. In (Y. Rossetti & A. Revonsuo, eds)

Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing. John Benjamins.

Schacter, D.L. 1992. Implicit knowledge: New perspectives on unconscious processes. Proceedings of the National Academy of Sciences USA 89:11113-17.

Schwartz, M. 1981. Criteria for physiological substrates of unconscious processes. American Psychologist 36:434-435.

Shevrin, H. & Dickman, S. 1980. The psychological unconscious: A necessary assumption for all psychological theory? American Psychologist 35:421-34.

Shevrin, H., Smith, W.H., Fitzler, D.E. 1971. Average evoked response and verbal correlates of unconscious mental processes. Psychophysiology 8:149-62.

Shevrin, H. & Fritzler, D.E. 1968. Visual evoked response correlates of unconscious mental processes. Science 161:295-298.

Spitz, H.H. 1993. The role of the unconscious in thinking and problem solving. Educational Psychology 13:229-244.

Spitz, H.H. 1995. Calendar calculating idiots savants and the smart unconscious. New Ideas in Psychology 13:167-182.

Stein, D.J. (ed) 1997. _Cognitive Science and the Unconscious_. American Psychiatric Press.

Suzuki, K. & Yamadori, A. 2000. Intact verbal description of letters with diminished awareness of their forms. Journal of Neurology, Neurosurgery & Psychiatry 68:782-786.

Underwood, G. & Bright, J.E.H. 1996. Cognition with and without awareness. In (G. Underwood, ed) _Implicit Cognition_. Oxford University Press.

Wachtel, P.L. 1987. "The unconscious" and unconscious processes. Canadian Psychology 28:107-108.

Yu, J. & Bellezza, F. 2000. Process dissociation as source monitoring. Journal of Experimental Psychology: Learning, Memory, & Cognition 26:1518-1533.

6.2d Visual Consciousness [see also 6.1b, 6.1c, 6.2b, 6.2e, 6.2j]

Bachmann, T. 1997. Visibility of brief images: The dual-process approach. Consciousness and Cognition 6:491-518.

Bachmann, T. 1998. Fast dynamics of visibility of brief visual images: The perceptual-retouch viewpoint. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness II_. MIT Press.

Baxt, N. 1982. On the time necessary for a visual impression to come into consciousness. Psychological Research 44:1-12.

Blackmore, S.J., Brelstaff, G., Nelson, K. & Troscianko, T. 1995. Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. Perception 24:1075-81.

Bridgeman, B., Hendry, D. & Stark, L. 1975. Failure to detect displacements of the visual world during saccadic eye movements. Vision Research 15:719-22.

Bridgeman, B. 2000. Interactions between vision for perception and vision for behavior. In (Y. Rossetti & A. Revonsuo, eds) _Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_. John Benjamins.

Cogan, A.I. 1995. Vision comes to mind. *Perception* 24:811-26.

Di Lollo, V., Enns, J., & Rensink, R. 2000. Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal Of Experimental Psychology-General* 129:481-507.

Dixon, M., Smilek, D., Cudahy, C., & Merikle, P. 2000. Five plus two equals yellow: Mental arithmetic in people with synaesthesia is not coloured by visual experience. *Nature* 406:365.

Durgin, F.H. 1995. On the filling in of the visual blind spot: Some rules of thumb. *Perception* 24:827-40.

Fernandez-Duque, D. & Thornton, I. 2000. Change detection without awareness: Do explicit reports underestimate the representation of change in the visual system? *Visual Cognition* 7:323-344.

Grimes, J. 1996. On the failure to detect changes in scenes across saccades. In (K. Akins, ed) *_Perception_*. Oxford University Press.

Irwin, D.E. 1991. Information integration across saccadic eye movements. *Cognitive Psychology* 23:420-56.

Iwasaki, S. 1993. Spatial attention and two modes of visual consciousness. *Cognition* 49:211-233.

Lachter, J., Durgin, F., & Washington, T. 2000. Disappearing percepts: Evidence for retention failure in metacontrast masking. *Visual Cognition* 7:269-279.

Mack, A. & Rock, I. 1998. *_Inattentional Blindness_*. MIT Press.

McConkie, G.W. & Zola, D. 1979. Is visual information integrated across successive fixations in reading? *Perception and Psychophysics* 25:221-24.

Natsoulas, T. 1993. An introduction to reflective seeing. *Journal of Mind and Behavior* 14:235-56.

Natsoulas, T. 1994. An introduction to reflective seeing. *Journal of Mind and Behavior* 15:351-74.

Nijhawan, R. & Khurana, B. 2000. Conscious registration of continuous and discrete visual events. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Noe, A., Pessoa, L., & Thompson, E. 2000. Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition* 7: 93-106.

O'Regan, J.K. 1992. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology* 46:461-88.

O'Regan, J., Deubel, H., Clark, J., & Rensink, R. 2000. Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition* 7:191-211.

Palmer, S. 1999. *_Vision Science: Photons to Phenomenology_*. MIT Press.

Pashler, H. 1988. Familiarity and visual change detection. *Perception and Psychophysics* 41:191-201.

Pessoa, L., Thompson, E. & Noe, A. 1998. Finding out about filling in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*.

Phillips, W.A. 1974. On the distinction between sensory storage and visual

short-term memory. Perception and Psychophysics 16:283-90.

Ramachandran, V.S. 1992. Filling in gaps in perception: Part I. Current Directions in Psychological Science 1:199-205.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. 1997. To see or not to see: The need for attention to perceive changes in scenes. Psychological Science 8:368-373.

Rensink, R., O'Regan, J., & Clark, J. 2000. On the failure to detect changes in scenes across brief interruptions. Visual Cognition Special 7:127-145.

Scott-Brown, K., Baker, M., & Orbach, H. 2000. Comparison blindness. Visual Cognition 7:253-267.

Shapiro, K. 2000. Change blindness: Theory or paradigm? Visual Cognition 7:83-91.

Shore, D., & Klein, R. 2000. The effects of scene inversion on change blindness. Journal of General Psychology 127:27-43.

Simons D.J. & Levin, D.T. 1997. Change blindness. Trends in Cognitive Science 1:241-82.

Simons, D. 2000. Current approaches to change blindness. Visual Cognition 7:1-15.

Sperling, G. 1960. The information available in visual presentations. Psychological Monographs 74:1-29.

Wallis, G. & Buelthoff, H. 2000. What's scene and not seen: Influences of movement and task upon what we see. Visual Cognition 7:175-190.

6.2e Consciousness and Attention [see also 6.2d]

Arvidson, P.S. 1996. Toward a phenomenology of attention. Human Studies 19:71-84.

Baars, B.J. 1997. Some essential differences between consciousness and attention, perception, and working memory. Consciousness and Cognition 6:363-371.

Baars, B.J. 1999. Attention vs. consciousness in the visual brain: Differences in conception, phenomenology, behavior, neuroanatomy, and physiology. Journal of General Psychology 126:224-33.

Baddeley, A.D. & Weiskrantz, L. (eds) 1993. _Attention: Selection, Awareness, and Control_. Oxford University Press.

Bridgeman, B. 1986. Relations between the physiology of attention and the physiology of consciousness. Psychological Research 48:259-266.

Cobb, C. 1955. Awareness, attention, and physiology of the brain stem. In (P. Hoch & J. Zubin, eds) _Experimental Psychopathology_.

Coslett, H.B. 1997. Consciousness and attention. Seminars in Neurology 17:137-44.

Cowan, N. & Wood, N.L. 1997. Constraints on awareness, attention, processing, and memory: Some recent investigations with ignored speech. Consciousness and Cognition 6:182-203.

Csikszentmihalyi, M. 1978. Attention and the holistic approach to behavior.

In (K. Pope & J. Singer, eds) *_The Stream of Consciousness: Scientific Investigation into the Flow of Experience_*. Plenum.

Hardcastle, V.G. 1997. Attention versus consciousness: A distinction with a difference. Manuscript.

He, S., Cavanagh, P. & Intriligator, J. 1996. Attentional resolution and the locus of visual awareness. *Nature* 383:334-37.

Hochberg, J. 1970. Attention, organization, and consciousness. In (D. Mostofsky, ed) *_Attention: Contemporary Theory and Analysis_*. Appleton-Century-Crofts.

Iwasaki, S. 1993. Spatial attention and two modes of visual consciousness. *Cognition* 49:211-233.

LaBerge, D. 1997. Attention, awareness, and the triangular circuit. *Consciousness and Cognition* 9:149-81.

LaBerge, D., Auclair, L., & Sieroff, E. 2000. Preparatory attention: Experiment and theory. *Consciousness & Cognition* 9:396-434.

Loper, A.B. & Hallahan, D.P. 1982. Meta-attention: The development of awareness of the attentional process. *Journal of General Psychology* 106:27-33.

Mack, A. & Rock, I. 1998. *_Inattentional Blindness_*. MIT Press.

Merikle, P.M. & Joordens, S. 1997. Parallels between perception without attention and perception without awareness. *Consciousness and Cognition* 6:219-36.

McCormick, P.A. 1997. Orienting attention without awareness. *Journal of Experimental Psychology: Human Perception & Performance* 23:168-180.

Newman, J.B., Baars, B.J., & Cho, S. 1997. A neural global workspace model for conscious attention. *Neural Networks* 10:1195-1206.

Newman, J.B. 1995. Thalamic contributions to attention and consciousness. *Consciousness and Cognition* 4:172-93.

Newsome, W.T. 1996. Visual attention: spotlights, highlights and visual awareness. *Current Biology* 6:357-60.

Posner, M.I. 1994. Attention: The mechanisms of consciousness. *Proceedings of the National Academy of Sciences USA* 91:7398-7403.

Posner, M.I. & Rothbart, M.K. 1992. Attentional mechanisms and conscious experience. In (A. Milner & M. Rugg, eds) *_The Neuropsychology of Consciousness_*. Academic Press.

Prinzmetal, W., Amiri, H., Allen, K. & Edwards, T. 1997. The phenomenology of attention, part 1: Color, location, orientation, and "clarity". *Journal of Experimental Psychology: Human Perception and Performance*.

Prinzmetal, W., Nwachuku, I., Bodanski, L., & Blumenfeld, L. 1997. The phenomenology of attention, part 2: Brightness and contrast. *Consciousness and Cognition* 6:372-412.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8:368-373.

Rensink, R. 2000. Visual search for change: A probe into the nature of attentional processing. *Visual Cognition* 7:345-376.

Scheier, M.F. Matthews, K.A. & Carver, C.S. 1983. Focus of attention and awareness of bodily states. In (G. Underwood, ed) *_Aspects of Consciousness, Volume 3: Awareness and Self-Awareness_*. Academic Press.

Scholl, B. 2000. Attenuated change blindness for exogenously attended items in a flicker paradigm. *Visual Cognition* 7:377-396.

Shiffrin, R.M. 1997. Attention, automatism, and consciousness. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Smilek, D., Eastwood, J., & Merikle, P. 2000. Does unattended information facilitate change detection? *Journal of Experimental Psychology: Human Perception & Performance* 26:480-487.

Umla, C. & Moscovitch, M. 1994. *_Attention and Performance 15: Conscious and Nonconscious Information Processing_*. MIT Press.

Underwood, G. 1977. Attention, awareness, and hemispheric differences in word recognition. *Neuropsychologia* 15:61-67.

Underwood, G. 1983. Selective attention and selective awareness of conscious processes. In (G. Underwood, ed) *_Aspects of Consciousness, Volume 3: Awareness and Self-Awareness_*. Academic Press.

6.2f Consciousness and Memory

Alkire, M.T., Haier, R.J., Fallon, J.H., & Barker, S.J. 1996. PET imaging of conscious and unconscious verbal memory. *Journal of Consciousness Studies* 3:448-62.

Allik, J. 2000. Available and accessible information in memory and vision. In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.

Andreasen, N. 2000. Is schizophrenia a disorder of memory or consciousness? In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.

Baddeley, A. 1993. Working memory and conscious awareness. In (A. Collins, S. Gathercole, M. Conway, & P. Morris, eds) *_Theories of Memory_*. Lawrence Erlbaum.

Balint, E. 1987. Memory and consciousness. *International Journal of Psychoanalysis* 68:475-483.

Barba, G. 2000. Memory, consciousness, and temporality: What is retrieved and who exactly is controlling the retrieval? In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.

Brainerd, C.J., Stein, L.M., & Reyna, V.F. 1998. On the development of conscious and unconscious memory. *Developmental Psychology* 34:342-357.

Brewer, W.F. 1992. Phenomenal experience in laboratory and autobiographical memory. In (M. Conway, D. Rubin, H. Spinnler, & W. Wagenaar, eds) *_Theoretical Perspectives on Autobiographical Memory_*. Kluwer.

Brewer, W.F. 1996. What is recollective memory? In (D. Rubin, ed) *_Remembering our Past: Studies in Autobiographical Memory_*. Cambridge University Press.

Buchner, A., Erdfelder, E., Vaterrodt-Plunnecke, B. 1995. Toward unbiased

measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General* 124:137-60.

Cavanaugh, J.C. 1989. The importance of awareness in memory aging. In (L. Poon, D. Rubin, & B. Wilson, eds) *_Everyday Cognition in Adulthood and Late Life_*. Cambridge University Press.

Clark, R.E. & Squire, L.R. 1998. Classical conditioning and brain systems: The role of awareness. *Science* 280:77-81.

Cloitre, M. 1997. Conscious and unconscious memory: A model of functional amnesia. In (D. Stein, ed) *_Cognitive Science and the Unconscious_*. American Psychiatric Press.

Conway, M.A. & Dewhurst, S.A. 1995. The self and recollective experience. *Applied Cognitive Psychology* 9:1-19.

Dalla Barba, G. 2000. Memory, consciousness, and the brain. *Brain & Cognition* 42:20-22.

Duezel, E. 2000. What brain activity tells us about conscious awareness of memory retrieval. In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.

Duzel E., Yonelinas A.P., Mangun G.R., Heinze H.J., & Tulving E. 1997. Event related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences of the United States of America* 94:5973-8.

Eich, E. 1984. Memory for unattended events: Remembering with and without awareness. *Memory and Cognition* 12:105-11.

Erdelyi, M. 1984. The recovery of unconscious (inaccessible) memories: Laboratory studies of hypermnnesia. In (G. Bower, ed) *_The Psychology of Learning and Motivation_*. Academic Press.

Gardiner, J.M. 1988. Functional aspects of recollective experience. *Memory and Cognition* 16:309-13.

Gardiner, J.M. & Parkin, A.J. 1990. Attention and recollective experience in recognition memory. *Memory and Cognition* 18:579-583.

Gardiner, J.M. 1993. Recognition memory and awareness: An experiential approach. *European Journal of Cognitive Psychology* 5:337-46.

Gardiner, J.M. 1996. On consciousness in relation to memory and learning. In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.

Gardiner, J.M., Ramponi, C. & Richardson-Klavehn, A. 1998. Experiences of remembering, knowing, and guessing. *Consciousness and Cognition* 7:1-26.

Gardiner, J. & Richardson-Klavehn, A. 2000. Remembering and knowing. In (E. Tulving & Craik, F., eds) *_The Oxford Handbook of Memory_*. Oxford University Press.

Gennaro, R.J. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333-47.

Gregg, V.H. & Gardiner, J.M. 1994. Recognition memory and awareness: A large effect of study-test modalities on "know" responses following a highly perceptual orienting task. *European Journal of Cognitive Psychology* 6:137-47.

Hamann, S.B. & Squire, L.R. 1997. Intact perceptual memory in the absence of conscious memory. *Behavioral Neuroscience* 111:850-54.

Hirshman, E. & Master, S. 1997. Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory and Cognition* 25:345-351.

Hirst, W. 1989. On consciousness, recall, recognition, and the architecture of memory. In (S. Lewandowsky, J. Dunn, & K. Kirsner, eds) *Implicit Memory: Theoretical Issues*. Lawrence Erlbaum.

Jacoby, L.L. & Witherspoon, D. 1982. Remembering without awareness. *Canadian Journal of Psychology* 36:300-324.

Jacoby, L.L. & Kelley, C.M. 1987. Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin* 13:314-36.

Jacoby, L.L. 1991. A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* 30:513-41.

Jacoby, L.L. & Kelley, C.M. 1991. Unconscious influences of memory: Dissociations and automaticity. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Jacoby, L.L., Toth, J.P. & Yonelinas, A.P. 1993. Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General* 122:139-54.

Jacoby, L.L., Toth, J.P., Yonelinas, A.P. & Debner, J.A. 1994. The relation between conscious and unconscious influences: Independence or redundancy? *Journal of Experimental Psychology: General*.

Jacoby, L.L., Yonelinas, A.P., & Jennings, J.M. 1997. The relation between conscious and unconscious (automatic) influences: A declaration of independence. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Johnson, M.K., Foley, M.A., Suengas, A.G. & Raye, C.L. 1988. Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General* 117:371-76.

Joordens, S. & Merikle, P.M. 1993. Independence or redundancy? Two models of conscious and unconscious influences. *Journal of Experimental Psychology: General* 122:462-67.

Kelley, C.M. & Jacoby, L.L. 1990. The construction of subjective experience: Memory attributions. *Mind and Language* 5:49-68.

Kelley, C.M. & Lindsay, D.S. 1996. Conscious and unconscious forms of memory. In (E. Bjork & R. Bjork, eds) *Memory: Handbook of Perception and Cognition*. Academic Press.

Kihlstrom, J.F. 1995. Memory and consciousness: An appreciation of Claparede and "Recognition et Moitte". *Consciousness and Cognition* 4:379-86.

Kuhlmann, F. 1996. On the analysis of the memory consciousness: A study in the mental imagery and memory of meaningless visual forms. *Psychological Review* 13:316-48.

Lampinen, J.M., Neuschatz, J.S., & Payne, D.G. 1998. Memory illusions and consciousness: Examining the phenomenology of true and false memories. *Current Psychology: Developmental, Learning, Personality, Social* 16:181-224.

- Larsen, S.F. 1998. What is it like to remember? On phenomenal qualities of memory. In (C. Thompson, J. Read, D. Bruce, D. Payne, & M. Toglia, eds) *_Autobiographical and Eyewitness Memory: Theoretical and Applied Perspectives_*. Lawrence Erlbaum.
- Light, L.L., Singh, A., & Capps, J.L. 1986. Dissociation of memory and awareness in young and older adults. *Journal of Clinical & Experimental Neuropsychology* 8:62-74.
- Lockhart, R.S. 1989. Consciousness and the function of remembered episodes. In (H. Roediger & F. Craik, eds) *_Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving_*.
- Mandler, G. 1989. Memory: Conscious and unconscious. In (P. Solomon, G. Goethals, C. Kelley, & R. Stephens, eds) *_Memory: Interdisciplinary Approaches_*. Springer-Verlag.
- Moscovitch, M. & Umiltà, C. 1991. Conscious and nonconscious aspects of memory: A neuropsychological framework of modules and central systems. In (R. Lister & H. Weingartner, eds) *_Perspectives on Cognitive Neuroscience_*. Oxford University Press
- Moscovitch, M. 1992. A neuropsychological model of memory and consciousness. In (L. Squire & N. Butters, eds) *_Neuropsychology of Memory_*. Guilford Press.
- Moscovitch, M., Goshen-Gottstein, Y. & Vriezen, E. 1994. Memory without conscious recollection: A tutorial review from a neuropsychological perspective. In (C. Umiltà and M. Moscovitch, eds) *_Consciousness and Unconscious Information Processing: Attention and Performance 15_*. MIT Press.
- Moscovitch, M. 1995. Recovered consciousness: A hypothesis concerning modularity and episodic memory. *Journal of Clinical and Experimental Neuropsychology*, 17:276-90.
- Moscovitch, M. 1995. Models of consciousness and memory. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.
- Moscovitch, M. 2000. Theories of memory and consciousness. In (E. Tulving & F. Craik, eds) *_The Oxford Handbook of Memory_*. Oxford University Press.
- Natsoulas, T. 1986. Consciousness and memory. *Journal of Mind and Behavior* 7:463-501.
- Oakhill, J. & Kyle, F. 2000. The relation between phonological awareness and working memory. *Journal of Experimental Child Psychology* 75:152-164.
- Paller, K.A., Kutas, M. & McIsaac, H.K. 1995. Monitoring conscious recollection via the electrical activity of the brain. *Psychological Science* 6:107-11.
- Paller, K. 2000. Neural measures of conscious and unconscious memory. *Behavioural Neurology* 12:127-141.
- Penfield, W. 1969. Consciousness, memory, and man's conditioned reflexes. In (H. Hyden, ed) *_On the Biology of Learning_*. Harcourt, Brace, and World.
- Rajaram, S. & Roediger, H.L. 1997. Remembering and knowing as states of consciousness during retrieval. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.
- Rajaram, S. 1998. The effects of conceptual salience and perceptual distinctiveness on conscious recollection. *Psychonomic Bulletin and Review* 5:71-78.

Reder, L.M. (ed) 1996. *_Implicit Memory and Metacognition_*. Lawrence Erlbaum.

Reingold, E.M. & Toth, P. 1996. Memory dissociations versus task dissociations: A controversy in progress. In (G. Underwood, ed) *_Implicit Cognition_*. Oxford University Press.

Richardson-Klavehn, A. & Gardner, J.M. 1996. Cross-modality priming in stem completion reflects conscious memory, but not voluntary memory. *Psychonomic Bulletin and Review* 3:238-44.

Richardson-Klavehn, A., Gardiner, J.M. & Java, R.I. 1996. Memory: Task dissociations, process dissociations and dissociations of consciousness. In (G. Underwood, ed) *_Implicit Cognition_*. Oxford University Press.

Roberts, H.M. 1971. Conscious experiences are a memory process. *Psychological Reports* 29:591-94.

Roediger, H.L. & Craik, F.I.M. (eds). *_Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving_*. Lawrence Erlbaum.

Ronnberg, J. & Archer, T. 1992. Purposive behaviour in cognition and perception: Considerations of awareness in memory. *Scandinavian Journal of Psychology* 33:86-91.

Rugg, M.D. 1995. Memory and consciousness: A selective review of issues and data. *Neuropsychologia* 33:1131-1141.

Schacter, D.L. 1987. Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:501-18.

Schacter, D.L. 1989. On the relation between memory and consciousness: Dissociable interactions and conscious experience. In (H. Roediger & F. Craik, eds) *_Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving_*.

Schacter, D.L., Bowers, J. & Booker, J. 1989. Intention, awareness, and implicit memory: The retrieval intentionality criterion. In (S. Lewandowsky, J. Dunn, & K. Kirsner, eds) *_Implicit Memory: Theoretical Issues_*. Lawrence Erlbaum.

Schacter, D.L. 1992. Consciousness and awareness in memory and amnesia: Critical issues. In (A. Milner & M. Rugg, eds) *_The Neuropsychology of Consciousness_*. Academic Press.

Schacter, D.L. 1995. Implicit memory: A new frontier for cognitive neuroscience. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.

Schacter, D.L. 1998. Memory and awareness. *Science* 280:59-60.

Stolz, J. & Merikle, P. 2000. Conscious and unconscious influences of memory: Temporal dynamics. *Memory* 8:333-343.

Toth, J.P., Lindsay, D.S. & Jacoby, L.L. 1992. Awareness, automaticity, and memory dissociations. In (L. Squire & N. Butters, eds) *_Neuropsychology of Memory_*. Guilford Press.

Toth, J.P. & Reingold, E.M. 1996. Beyond perception: Conceptual contributions to unconscious influences of memory. In (G. Underwood, ed) *_Implicit Cognition_*. Oxford University Press.

Tulving, E. 1985. Memory and consciousness. *Canadian Psychology* 26:1-12.

Tulving, E. 1987. Multiple memory systems and consciousness. *Human Neurobiology* 6:67-80.

Tulving, E. 1993. Varieties of consciousness and levels of awareness in memory. In (A. Baddeley & L. Weiskrantz, eds) *Attention: Selection, Awareness, and Control*. Oxford University Press.

Underwood, G. 1979. Memory systems and conscious processes. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 1*. Academic Press.

Verfaellie, M. & Keane, M.M. 1997. The neural basis of aware and unaware forms of memory. *Seminars in Neurology* 17:153-61.

Wheeler, M.A. Stuss, D.T. & Tulving, E. 1997. Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin* 121:331-54.

Wippich, W. 1992. Implicit and explicit memory without awareness. *Psychological Research* 54:212-24.

6.2g Consciousness and Learning

Berry, D.C. 1994. Implicit learning: Twenty-five years on. A tutorial. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance 15*. MIT Press.

Berry, D.C. 1997. *How Implicit is Implicit Learning?*. Oxford University Press.

Berry, D.C. & Dienes, Z. (eds) 1993. *Implicit Learning: Theoretical and Empirical Issues*. Lawrence Erlbaum Associates.

Brody, N. 1989. Unconscious learning of rules: Comment on Reber's analysis of implicit learning. *Journal of Experimental Psychology: General* 118:236-238.

Carlson, R.A. & Dulany, D.E. 1985. Conscious attention and abstraction in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11:45-58.

Cleeremans, A. 1993. Mechanisms of implicit learning: Connectionist models of sequence processing.

Conway, M.A., Collins, A.F., Anderson, S.J., & Cohen, G. 1998. Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General*.

Dienes, Z. & Berry, D. 1997. Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review* 4:3-23.

Dulany, D.E. 1968. Awareness, rules, and propositional control: A confrontation with S-R behavior theory. In (T. Dixon & D. Horton, eds) *Verbal Behavior and General Behavior Theory*. Prentice-Hall.

Dulany, D.E., Carlson, R.A. & Dewey, G.I. 1984. A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General* 113:541-555.

Dulany, D.E., Carlson, R.A., Dewey, G.I. 1985. On consciousness in syntactic learning and judgment: A reply to Reber, Allen, and Regan. *Journal of Experimental Psychology: General* 114:25-32.

Furedy, J., Damke, B., & Boucsein, W. 2000. Revisiting the learning-without-awareness question in human Pavlovian autonomic conditioning: Focus on extinction in a dichotic listening paradigm. *Integrative Physiological & Behavioral Science* 35:17-34.

- Gardiner, J.M. 1996. On consciousness in relation to memory and learning. In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.
- Kihlstrom, J.F. 1996. Perception without awareness of what is perceived, learning without awareness of what is learned. In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.
- Manns, J., Clark, R., & Squire, L. 2000. Awareness predicts the magnitude of single-cue trace eyeblink conditioning. *Hippocampus* 10:181-186.
- Marton, F. & Booth, S.A. 1997. *_Learning and Awareness_*. Lawrence Erlbaum.
- Neal, A. & Hesketh, B. 1997. Episodic knowledge and implicit learning. *Psychonomic Bulletin and Review* 4:24-37.
- Neal, A. & Hesketh, B. 1997. Future directions for implicit learning: Toward a clarification of issues associated with knowledge representation and consciousness. *Psychonomic Bulletin and Review* 4:73-78.
- Perruchet, P., Gallego, J. & Savy, I. 1990. A critical reappraisal of the evidence for unconscious abstraction of deterministic rules in complex experimental situations. *Cognitive Psychology* 22:493-516.
- Perruchet, P. & Pacteau, C. 1990. Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge. *Journal of Experimental Psychology: General* 119:264-75.
- Perruchet, P., Vinter, A., & Gallego, J. 1997. Implicit learning shapes new conscious percepts and representations. *Psychonomic Bulletin & Review* 4:43-48.
- Reber, A.S. 1967. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* 6:855-863.
- Reber, A.S. 1989. Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* 118:219-35.
- Reber, A.S. 1993. *_Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious_*. Oxford University Press.
- Reber, A.S. 1997. How to differentiate implicit and explicit modes of acquisition. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.
- Reber, A.S., Allen, R., & Regan, S. 1985. Syntactical learning and judgment, still unconscious and still abstract: Comment on Dulany, Carlson, and Dewey. *Journal of Experimental Psychology: General* 114:17-24.
- Shanks, D.R., Green, R.E.A., & Kolodny, J.A. 1994. A critical examination of the evidence for unconscious (implicit) learning. In (C. Umiltà and M. Moscovitch, eds) *_Consciousness and Unconscious Information Processing: Attention and Performance 15_*. MIT Press.
- Shanks, D.R. & St. John, M.F. 1994. Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* 17:367-447.
- Sno, H. 2000. Deja vu and jamais vu. In (G. Berrios & J. Hodges, eds) *_Memory Disorders in Psychiatric Practice_*. Cambridge University Press.
- Stadler, M.A. & Roediger, H.L. 1998. The question of awareness in research on implicit learning. In (M. Stadler & P. Frensch, eds) *_Handbook of Implicit Learning_*. Sage Publications
- Stadler, M.A., & Frensch, P.A. 1998. *_Handbook of Implicit Learning_*. Sage

Publications.

Whittlesea, B.W.A. & Dorken, M.D. 1997. Implicit learning: Indirect, not unconscious. *Psychonomic Bulletin & Review* 4:63-67.

Wong, P.S., Bernat, E., Bunce, S., & Shevrin, H. 1997. Brain indices of nonconscious associative learning. *Consciousness and Cognition* 6:519-544.

6.2h Consciousness and Metacognition

Brown, R. & McNeill, D. 1966. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior* 5:325-37.

Brown, S. 2000. Tip-of-the-tongue phenomena: An introductory phenomenological analysis. *Consciousness & Cognition* 9:516-537.

Hart, J.T. 1965. Memory and the feeling-of-knowing experience. *Journal of Educational Psychology* 56:208-16.

Johnson, M.K. 1988. Reality monitoring: An experimental phenomenological approach. *Journal of Experimental Psychology: General* 117:390-94.

Johnson, M.K. 1991. Reflection, reality monitoring, and the self. In (R. Kunzendorf, ed) *_Mental Imagery_*. Plenum Press.

Johnson, M.K. & Reeder, J.A. 1997. Consciousness as meta-processing. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Kahan, T.L. & LaBerge, S. 1994. Lucid dreaming as metacognition: Implications for cognitive science. *Consciousness & Cognition* 3:246-64.

Mangan, B. 2000. What feeling is the "feeling of knowing?" *Consciousness & Cognition* 9:538-544.

Metcalfe, J. & Shimamura, P. 1994. *_Metacognition: Knowing about Knowing_*. MIT Press.

Nelson, T.O. 1996. Consciousness and metacognition. *American Psychologist* 51:102-16.

Nelson, T.O. 1992. *_Metacognition: Core Readings_*. Allyn and Bacon.

Nisbett, R. & Wilson, T. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review* 84:231-59.

Otani, H. & Hodge, M. 1991. mechanisms of feelings of knowing: The role of elaboration and familiarity. *Psychological Record* 41:523-35.

Reder, L.M. (ed) 1996. *_Implicit Memory and Metacognition_*. Lawrence Erlbaum.

Reder, L.M. & Schunn, C.D. 1996. Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In (L. Reder, ed) *_Implicit Memory and Metacognition_*. Lawrence Erlbaum.

Ricciardelli, L.A. 1993. Two components of metalinguistic awareness: Control of linguistic processing and analysis of linguistic knowledge. *Applied Psycholinguistics* 14:349-367.

Rosenthal, D.M. 1998. Consciousness and metacognition. In (D. Sperber, ed) *_Metarepresentation_*. Oxford University Press.

Wegner, D.M. 1989. *_White Bears and Other Unwanted Thoughts: Suppression,*

Obsession, and the Psychology of Mental Control_. Penguin.

Wegner, D.M. 1997. Why the mind wanders. In (J. Cohen & J. Schooler, eds) Scientific Approaches to Consciousness. Lawrence Erlbaum.

White, P. 1980. Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review* 87:105-12.

White, P. 1983. Knowing our own minds: Conscious awareness and verbal reports. In (G. Underwood, ed) Aspects of Consciousness, Volume 3: Awareness and Self-Awareness. Academic Press.

Wilson, T.D. 1997. The psychology of metapsychology. In (J. Cohen & J. Schooler, eds) Scientific Approaches to Consciousness. Lawrence Erlbaum.

6.2i Consciousness and Control

Baars, B.J. 1987. What is conscious in the control of action? A modern ideomotor theory of voluntary action. In (D. Gorfein & R. Hoffman, eds) Learning and Memory: The Ebbinghaus Centennial Symposium. Lawrence Erlbaum.

Baars, B.J. 1992. Experimental Slips and Human Error: Exploring the Architecture of Volition. Plenum Press.

Baars, B.J. 1993. Why volition is a foundation issue for psychology. *Consciousness and Cognition* 2:281-309.

Bargh, J.A. 1996. Automaticity in action: The unconscious as repository of chronic goals and motives. In (P. Gollwitzer & J. Bargh, eds) The Psychology of Action: Linking Cognition and Motivation to Behavior. Guilford.

Bargh, J.A. 1996. Automaticity in social psychology. In (E. Higgins & A. Kruglanski, eds) Social Psychology: Handbook of Basic Principles. Guilford.

Bargh, J.A. 1994. The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In (R. Wyer & T. Srull, eds) Handbook of Social Cognition. Lawrence Erlbaum.

Bayles, G.H. & Cleary, P.J. 1986. The role of awareness in the control of frontalis muscle activity. *Biological Psychology* 22:23-35.

Carr, T.H., McCauley, C., Sperber, R.D., & Parmelee, C.M. 1982. Words, pictures, and priming: On semantic activation, conscious identification, and the automaticity of information processing. *Journal of Experimental Psychology: Human Perception & Performance* 8:757-777.

Daprati, E., Franck, N., Georgieff, N.,; Proust, J. 1997. Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition* 65:71-86.

Delabarre, E.B. 1911. Volition and motor consciousness: Theory. *Psychological Bulletin* 8:378-82.

Delabarre, E.B. 1912. Volition and motor consciousness: Theory. *Psychological Bulletin* 9:409-13.

Delabarre, E.B. 1913. Volition and motor consciousness. *Psychological Bulletin* 10:441-44.

Dewan, E.M. 1976. Consciousness as an emergent causal agent in the context of control system theory. In (G. Globus, G. Maxwell, & I. Savodnik, eds) Consciousness and the Brain. Plenum Press.

- Gott, P.S., Hughes, E.C. & Whipple, K. 1984. Voluntary control of two lateralized conscious states: Validation of electrical and behavioral studies. *Neuropsychologia* 22:65-72.
- Gordon, A.M. & Rosenbaum, D.A. 1984. Conscious and subconscious arm movements: Application of signal detection theory to motor control. *Bulletin of the Psychonomic Society* 22:214-216.
- Gray, J.A. 1998. Abnormal contents of consciousness: The transition from automatic to controlled processing. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.
- Hilgard, E.R. 1979. Consciousness and control: Lessons from hypnosis. *Australian Journal of Clinical & Experimental Hypnosis* 7:103-15.
- Horowitz, M.J. & Stinson, C.H. 1995. Consciousness and processes of control. *Journal of Psychotherapy Practice and Research* 4:123-139.
- Jacoby, L.L. 1991. A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* 30:513-41.
- Jacoby, L.L., Ste-Marie, D. & Toth, J.P. 1993. Redefining automaticity: Unconscious influences, awareness, and control. In (A. Baddeley & L. Weiskrantz, eds) *_Attention: Selection, Awareness, and Control_*. Oxford University Press.
- Kamiya, J. 1968. Conscious control of brain waves. *Psychology Today* 1:56-60.
- Kimble, G.A. & Perlmutter, L.C. 1970. The problem of volition. *Psychological Review* 77:361-84.
- Langer, E.J. 1992. Matters of mind: Mindfulness/mindlessness in perspective. *Consciousness and Cognition* 1:289-305.
- Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8:529-566.
- Oswald, M. & Gadenne, V. 2000. Are controlled processes conscious? In (W. Perrig & A. Grob, eds) *_Control of Human Behavior, Mental Processes, and Consciousness: Essays in Honor of the 60th Birthday of August Flammer_*. Lawrence Erlbaum Associates.
- Plotkin, W.B. 1976. On the self-regulation of the occipital alpha rhythm: Control strategies, states of consciousness, and the role of physiological feedback. *Journal of Experimental Psychology: General* 105:66-99.
- Plotkin, William B. 1981. A rapprochement of the operant-conditioning and awareness views of biofeedback training: The role of discrimination in voluntary control. *Journal of Experimental Psychology: General* 110:415-428.
- Posner, M. & Snyder, C.R.R. 1975. Attention and cognitive control. In (R. Solso, ed) *_Information Processing and Cognition: The Loyola Symposium_*. Lawrence Erlbaum.
- Raichle, M.E. 1997. Automaticity: From reflective to reflexive information processing. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *_Cognition, Computation, and Consciousness_*. Oxford University Press.
- Schneider, W.E. & Shiffrin, R.M. 1977. Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* 84:1-66.

Shiffrin, R.M. & Schneider, W.E. 1977. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review* 84:128-90.

Tzelgov, J. 1997. Specifying the relations between automaticity and consciousness: A theoretical note. *Consciousness and Cognition* 6:441-51.

Tzelgov, J., Porat, Z. & Henik, A. 1997. Automaticity and consciousness: Is perceiving the word necessary for reading it? *American Journal of Psychology* 110:429-48.

Tzelgov, J. 1997. Automatic but conscious: That is how we act most of the time. In (R. Wyer, ed) *The Automaticity of Everyday Life*. Lawrence Erlbaum.

Uleman, J.S. 1987. Consciousness and control: The case of spontaneous trait inferences. *Personality and Social Psychology Bulletin* 13:337-54.

Umiltà, C. 1988. The control operations of consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

White, W.A. 1920. Extending the field of conscious control. *Mental Hygiene* 4:857-66.

Zelazo, P.D. & Frye, D. 1997. Cognitive complexity and control: A theory of the development of deliberate reasoning and intentional action. In (M. Stamenov, ed) *Language Structure, Discourse, and the Access to Consciousness*. John Benjamins.

6.2j Consciousness and Imagery

Ahsen, A. 1991. Imagery and consciousness: Putting together poetic, mythic and social realities. *Journal of Mental Imagery* 15:63-97.

Ahsen, A. 1991. A second report on AA-VVIQ: Role of vivid and unvivid images in consciousness research. *Journal of Mental Imagery* 15:1-31.

Ahsen, A. 1993. Imagery paradigm: Imaginative consciousness in the experimental and clinical setting. *Journal of Mental Imagery*.

Arnheim, R. 1994. Consciousness: An island of images. *Journal of Theoretical and Philosophical Psychology* 14:121-27.

Bichowsky, F.R. 1926. The mechanism of consciousness: Images. *American Journal of Psychology* 37:557-564.

Frick, R.W. 1987. A dissociation of conscious visual imagery and visual short-term memory. *Neuropsychologia* 25:707-12.

Hampson, P.J. & Morris, P.E. 1990. Imagery, consciousness, and cognitive control: The BOSS model reviewed. In (P. Hampson, D. Marks, & J. Richardson, Eds) *Imagery: Current Developments*. Routledge.

Hebb, D.O. 1968. Concerning imagery. *Psychological Review* 75:466-77.

Horne, P.V. 1993. The nature of imagery. *Consciousness and Cognition* 2:58-82.

Hubbard, T.L. 1996. The importance of a consideration of qualia to imagery and cognition. *Consciousness and Cognition* 3:327-58.

Ishai, A. & Sagi, D. 1998. Visual imagery and visual perception: The role of memory and conscious awareness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Krellenstein, M.F. 1995. Unsolvability problems, visual imagery, and explanatory satisfaction. *Journal of Mind and Behavior* 16:235-54.

Kunzendorf, R.G. 1990. The causal efficacy of consciousness in general, imagery in particular: A materialist perspective. In (R. Kunzendorf, ed) *Mental Imagery*. Plenum Press.

Kunzendorf, R.G., Justice, M., & Capone, D. 1997. Conscious images as "centrally excited sensations": A developmental study of imaginal influences on the ERG. *Journal of Mental Imagery* 21:155-66.

Lehmann, D., Henggler, B., Koukkan, M. & Michel, M. 1993. Source localization of brain electric field frequency bands during conscious, spontaneous visual imagery and abstract thought. *Cognitive Brain Research* 1:203-20.

Mandler, G. 1984. Consciousness, imagery, and emotion -- with special reference to autonomic imagery. *Journal of Mental Imagery* 8:87-94.

Marks, D.F. 1977. Imagery and consciousness: A theoretical review from an individual differences perspective. *Journal of Mental Imagery* 1:275-90.

Marks, D.F. 1983. Imagery and consciousness: A theoretical review. In (A. Sheikh, ed) *Imagery: Current Theory, Research, and Application*. Wiley.

Marks, D.F. 1990. On the relationship between imagery, body, and mind. In (P. Hampson, D. Marks, & J. Richardson, eds) *Imagery: Current Developments*. Routledge.

Mavromatis, A. 1987. On shared states of consciousness and objective imagery. *Journal of Mental Imagery* 11:125-30.

Morris, P.E. & Hampson, P.J. 1983. *Imagery and Consciousness*. Academic Press.

Newton, N. 1982. Experience and imagery. *Southern Journal of Philosophy* 21:475-87.

Richardson, A. 2000. Individual differences in visual imagination imagery. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Sheehan, P.W. & Lewis, S.E. 1974. Subjects' reports of confusion in consciousness and the arousal of imagery. *Perceptual and Motor Skills* 38:731-34.

6.2k Consciousness and Emotion

Cioffi, D. 1991. Sensory awareness versus sensory impression: Affect and attention interact to produce somatic meaning. *Cognition and Emotion* 5:275-94.

Damasio, A. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.

DeLancey, C. 1996. Emotion and the function of consciousness. *Journal of Consciousness Studies* 3:492-99.

Dimberg, U., Thunberg, M., & Elmehed, K. 2000. Unconscious facial reactions to emotional facial expressions. *Psychological Science* 11:86-89.

Ellis, R.D. 1995. *Questioning Consciousness: The Interplay of Imagery, Cognition, and Emotion in the Human Brain*. John Benjamins.

Ellis, R. 2000. *The Caldron of Consciousness: Motivation, Affect and*

Self-organization_. John Benjamins.

Forgas, J. & Ciarrochi, J. 2000. Affect infusion and affect control: The interactive role of conscious and unconscious processing strategies in mood management. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

Gray, J.A. 1999. Cognition, emotion, conscious experience and the brain. In (T. Dalgleish & M.J. Powers, eds) *_Handbook of Cognition and Emotion_*. Wiley.

Heilman, K. 2000. Emotional experience: A neurological model. In (R. Lane, L. Nadel, & G. Ahern, J. Allen, & A. Kaszniak, eds) *_Cognitive Neuroscience of Emotion_*. Oxford University Press.

Katz, J.M. 1983. Altered states of consciousness and emotion. *Imagination, Cognition and Personality* 2:37-50.

Kunst-Wilson, W.R. & Zajonc, R.B. 1980. Affective discrimination of stimuli that cannot be recognized. *Science* 207:557-58.

Lane, R.D. Reiman, E., Ahern, G., Schwartz, G.E., & Yun, L. 1998. Anterior cingulate cortex participates in the conscious experience of emotion. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Lane, R. 2000. Neural correlates of conscious emotional experience. In (R. Lane, L. Nadel, & G. Ahern, J. Allen, & A. Kaszniak, eds) *_Cognitive Neuroscience of Emotion_*. Oxford University Press.

Lane, R.D., Ahern, G.L., Schwartz, G.E. & Kaszniak, A.W. 1997. Is alexithymia the emotional equivalent of blindsight? *Biological Psychiatry* 42:834-44.

Ledoux, J.E. 1995. In search of an emotional system in the brain: Leaping from fear to emotion and consciousness. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.

Lewis, M. 1997. The self in self-conscious emotions. In (J. Snodgrass, R. Thompson, eds) *_The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept_*. New York Academy of Sciences.

Lishman, W.A. 1971. Emotion, consciousness, and will after brain bisection in man. *Cortex* 7:181-92.

Matsumoto, D. & Lee, M. 1993. Consciousness, volition, and the neuropsychology of facial expressions of emotion. *Consciousness and Cognition* 2:237-54.

Morris, J.S., Ohman, A., & Dolan, R.J. 1998. Conscious and unconscious emotional learning in the human amygdala. *Nature* 393:467-470.

Niedenthal, P.M. 1990. Implicit perception of affective information. *Journal of Experimental Social Psychology* 26:505-27.

Ochsner, K. 2000. Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology: General* 129:242-261.

Ohman, A., Flykt, A., & Lundqvist, D. 2000. Unconscious emotion: Evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. In (R. Lane, L. Nadel, & G. Ahern, eds) *_Cognitive Neuroscience of Emotion_*. Oxford University Press.

Peper, M. 2000. Awareness of emotions: A neuropsychological perspective. In

(R. Ellis, ed) *_The Caldron of Consciousness: Motivation, Affect and Self-organization_*. John Benjamins Publishing Company.

Pratto, F. 1994. Consciousness and automatic evaluation. In (P. Niedenthal & S. Kitayama, eds) *_The Heart's Eye: Emotional Influences in Perception and Attention_*. Academic Press.

Reiman, E.M., Lane, R.D., Ahern, G.L., Schwartz, G.E. 1996. Positron emission tomography, emotion, and consciousness. In (S. Hamreoff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Rolls, E.T. 1995. A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In (M. Gazzaniga, eds) *_The Cognitive Neurosciences_*. MIT Press.

Safran, J.D. & Greenberg, L.S. 1987. Affect and the unconscious: A cognitive perspective. In (R. Stern, ed) *_Theories of the Unconscious and Theories of the Self_*. Analytic Press.

Wakefield, J.C. 1991. Why emotions can't be unconscious: An exploration of Freud's essentialism. *Psychoanalysis and Contemporary Thought* 14:29-67.

Weiskrantz, L. 2000. Blindsight: Implications for the conscious experience of emotion. In (R. Lane, L. Nadel, & G. Ahern, eds) *_Cognitive Neuroscience of Emotion_*. Oxford University Press.

Zajonc, R. 2000. Feeling and thinking: Closing the debate over the independence of affect. In (J. Forgas, ed) *_Feeling and Thinking: The Role of Affect in Social Cognition_*. Cambridge University Press.

6.21 Consciousness, Sleep, and Dreaming -----

Arden, J.B. 1996. *_Consciousness, Dreams, and Self: A Transdisciplinary Approach_*. Psychosocial Press.

Bentley, E. 2000. *_Awareness: Biorhythms, Sleep and Dreaming_*. Routledge.

Bosinelli, M. 1995. Mind and consciousness during sleep. *Behavioural Brain Research* 69:195-201.

Broughton, R.J. 1982. Human consciousness and sleep/waking rhythms: A review and some neuropsychological considerations. *Journal of Clinical Neuropsychology* 4:193-218.

Combs, A. & Krippner, S. 1998. Dream sleep and waking reality: A dynamical view. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Flanagan, O. 1997. Prospects for a unified theory of consciousness or, what dreams are made of. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Foulkes, D. 1990. Dreaming and consciousness. *European Journal of Cognitive Psychology* 2:39-55.

Gackenbach, J. & LaBerge, S. 1988. *_Conscious Mind, Sleeping Brain: Perspectives on Lucid Dreaming_*. Plenum Press.

Green, C. & McGreery, C. 1994. *_Lucid Dreaming: The Paradox of Consciousness During Sleep_*. Routledge.

Hearne, K.M. 1992. Pro-lucid dreaming, lucid dreams, and consciousness. *Journal of Mental Imagery* 16:119-123.

- Hobson, J.A. & Strickgold, R. 1995. The conscious state paradigm: A neurocognitive approach to waking, sleeping, and dreaming. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.
- Hobson, J.A. 1998. The conscious state paradigm: A neuropsychological analysis of waking, sleeping, and dreaming. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.
- Hobson, J., Pace-Schott, E., & Stickgold, R. 2000. Consciousness: Its vicissitudes in waking and sleep. In (M. Gazzaniga, ed) *_The New Cognitive Neurosciences: 2nd Edition_*. MIT Press.
- Jones, B.E. 1998. The neural basis of consciousness across the sleep-waking cycle. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.
- Kahan, T.L. & LaBerge, S. 1994. Lucid dreaming as metacognition: Implications for cognitive science. *Consciousness & Cognition* 3:246-264.
- Kahan, T.L. & LaBerge, S. 1996. Cognition and metacognition in dreaming and waking: Comparisons of first and third-person ratings. *Dreaming* 6:235-249.
- Kahn, D., Pace-Schott, E.F. & Hobson, J.A. 1997. Consciousness in waking and dreaming: The roles of neuronal oscillation and neuromodulation in determining similarities and differences. *Neuroscience* 78:13-38.
- Khan, D., Krippner, S., & Combs, A. 2000. Dreaming and the self-organizing brain. *Journal of Consciousness Studies* 7:4-11.
- King, C.D. 1947. Dream and the problem of consciousness. *Journal of General Psychology* 37:15-24.
- Kleitman, N. 1957. Sleep, wakefulness, and consciousness. *Psychological Bulletin* 54:354-359.
- LaBerge, S. 1985. *_Lucid Dreaming_*. J.P. Tarcher.
- LaBerge, S., Levitan, L., & Dement, W.C. 1986. Lucid dreaming: Physiological correlates of consciousness during REM sleep. *Journal of Mind and Behavior* 7:251-258.
- LaBerge, S. 1990. Lucid dreaming: Psychophysiological studies of consciousness during REM sleep. In (R. Bootsen, J. Kihlstrom, & D. Schacter, eds) *_Sleep and Cognition_*. American Psychological Association Press.
- LaBerge, S. 1998. Dreaming and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.
- LaBerge, S. & DeGracia, D. 2000. Varieties of lucid dreaming experience. In (R. Kunzendorf & B. Wallace, eds) *_Individual Differences in Conscious Experience_*. John Benjamins.
- Lindsley, D.B. 1960. Attention, consciousness, sleep, and wakefulness. In (H. Magoun & V. Hall, eds) *_Handbook of Physiology. Section I: Neurophysiology_*. American Physiological Society.
- Makeig, S., Jung, T., & Sejnowski, T. 2000. Awareness during drowsiness: Dynamics and electrophysiological correlates. *Canadian Journal of Experimental Psychology* 54:266-273.
- Monnier, M. 1952. Experimental work on sleep and other variations of consciousness. In (H. Abramson, ed) *_Problems of Consciousness: Transactions*

of the Third Conference_. Josiah Macy Foundation.

Moorcroft, W. & Breitenstein, J. 2000. Awareness of time during sleep. *Annals of Medicine* 32:236-238.

Munglani, R. & Jones, J.G. 1992. Sleep and general anaesthesia as altered states of consciousness. *Journal of Psychopharmacology* 6:399-409.

Pare, D. & Llinas, R. 1995. Conscious and pre-conscious processes as seen from the standpoint of sleep-waking cycle neurophysiology. *Neuropsychologia* 33:1155-1168.

Revonsuo, A. 1995. Consciousness, dreams and virtual realities. *Philosophical Psychology* 8:35-58.

Simon, C.W. & Emmons, W. 1956. EEG, consciousness, and sleep. *Science* 124:1066-1069.

Stoyva, J. & Kamiya, J. 1968. Electrophysiological studies of dreaming as the prototype of a new strategy in the study of consciousness. *Psychological Review* 75:192-205.

6.2m Consciousness in Psychoanalysis

Balint, E. 1987. Memory and consciousness. *International Journal of Psychoanalysis* 68:475-483.

Barr, H. & Langs, R. 1972. The psychoanalytic theory of consciousness. In *_LSD: Personality and experience_*. Wiley-Interscience.

Bouveresse, J. 1995. *_Wittgenstein Reads Freud: The Myth of the Unconscious_*. Princeton University Press.

Bower, G.H. 1990. Awareness, the unconscious, and repression: An experimental psychologist's perspective. In (J. Singer, ed) *_Repression and Dissociation_*. University of Chicago Press.

Brakel L.W. 1989. Negative hallucinations, other irretrievable experiences and two functions of consciousness. *International Journal of Psychoanalysis* 70:461-89.

Burston, D. 1986. The cognitive and dynamic unconscious: A critical and historical perspective. *Contemporary Psychoanalysis* 22:133-57.

Chang, S.C. 1978. The psychology of consciousness. *American Journal of Psychotherapy* 32:105-116.

Curtis, R. 1992. A process view of consciousness and the "self": Integrating a sense of connectedness with a sense of agency. *Psychological Inquiry*, 3:29-32.

Eagle, M.N. 1987. The psychoanalytic and the cognitive unconscious. In (R. Stern, ed) *_Theories of the Unconscious and Theories of the Self_*. Analytic Press.

Epstein, S. 1994. Integration of the cognitive and the psychodynamic unconscious. *American Psychologist* 49:409-24.

Erdelyi, M.H. 1988. Issues in the study of unconscious and defense processes. In (M. Horowitz, ed) *_Psychodynamics and Cognition_*. University of Chicago Press.

Foulkes, D. 1964. Theories of dream formation and recent studies of sleep consciousness. *Psychological Bulletin* 62:236-47.

- Globus, G.G. 1974. The problem of consciousness. *Psychoanalysis and Contemporary Science* 3:40-69.
- Haldane, J. 1988. Psychoanalysis, cognitive psychology and self-consciousness. In (P. Clark & C. Wright, eds) *Mind, Psychoanalysis and Science*. Blackwell.
- Herzog, P.S. 1991. *Conscious and Unconscious: Freud's Dynamic Distinction Reconsidered*. International Universities Press.
- Horowitz, M.J. & Stinson, C.H. 1995. Consciousness and processes of control. *Journal of Psychotherapy Practice and Research* 4:123-139.
- Joseph E.D. 1987. The consciousness of being conscious. *Journal of the American Psychoanalytic Association* 35:5-22.
- Klein, G. 1959. Consciousness in psychoanalytic theory. *Journal of the American Psychoanalytic Association* 7:5-34.
- Kubie, L.S. 1954. Psychiatric and psychoanalytic considerations of the problem of consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.
- Levy, D. 1996. *Freud among the Philosophers: The Psychoanalytic Unconscious and its Philosophical Critics*. Yale University Press.
- Masek, R. 1989. The overlooked problem of consciousness in psychoanalysis: Pierre Janet revisited. *Humanistic Psychologist* 17:274-279.
- MacIntyre, A.C. 1958. *The Unconscious: A Conceptual Study*. London.
- Miller, L. 1997. Freud and consciousness: The first one hundred years of neuropsychodynamics in theory and clinical practice. *Seminars in Neurology* 17:171-77.
- Moraglia, G. 1991. The unconscious in information processing and analytical psychology. *Journal of Analytical Psychology* 36:27-36.
- Natsoulas, T. 1992. Toward an improved understanding of Sigmund Freud's conception of consciousness. *Journal of Mind & Behavior* 13:171-92.
- Natsoulas, T. 1984-1996. Freud and consciousness I-XI. *Psychoanalysis and Contemporary Thought* 7:195-232, 8:183-220, 12:97-123, 12:619-62, 14:69-108, 15:305-48, 16:67-101, 16:597-631, 19:461-94.
- Natsoulas, T. 1995. A rediscovery of Sigmund Freud. *Consciousness and Cognition* 4:300-322.
- Olds, D.D. 1992. Consciousness: A brain-centered, informational approach. *Psychoanalytic Inquiry* 12:419-44.
- Opatow, B. 1997. The real unconscious: Psychoanalysis as a theory of consciousness. *Journal of the American Psychoanalytic Association* 45:865-90.
- Orbach, I. 1995. *The Hidden Mind: Psychology, Psychotherapy, and Unconscious Processes*. Wiley.
- Piaget, J. 1973. The affective unconscious and the cognitive unconscious. *Journal of the American Psychoanalytic Association* 21:249-261.
- Power, M.J. & Brewin, C.R. 1991. From Freud to cognitive science: A contemporary account of the unconscious. *British Journal of Clinical Psychology* 30:289-310.

Power, M.J. 1997. Conscious and unconscious representations of meaning. In (M. Power & C. Brewin, eds) *The Transformation of Meaning in Psychological Therapies: Integrating Theory and Practice*. John Wiley.

Rapaport, D. 1951. Consciousness: A Psychopathological and psychodynamic view. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Second Conference*. Josiah Macy Foundation.

Rosenblatt A.D. & Thickstun J.T. 1994. Intuition and consciousness. *Psychoanalytic Quarterly* 63:696-714.

Rubinfine, D.L. 1973. Notes toward a theory of consciousness. *International Journal of Psychoanalytic Psychotherapy* 2:391-410.

Ryle, A. 1994. Consciousness and psychotherapy. *British Journal of Medical Psychology* 67:115-23.

Schimek, J.G. 1975. A critical re-examination of Freud's concept of unconscious mental representation. *International Review of Psychoanalysis* 2:171-87.

Shevrin, H. 1990. Subliminal perception and repression. In (J. Singer, ed) *Repression and Dissociation: Implications for Personality Theory, Psychopathology, and Health*. University of Chicago Press.

Shevrin, H. 1992. The Freudian unconscious and the cognitive unconscious: Identical or fraternal twins? In (J. Barron, M. Eagle, & D. Wolitzky, eds) *Interface of Psychoanalysis and Psychology*. American Psychological Association.

Shevrin, H., Williams, W.J., Marshall, R.E., & Brakel, L.A. 1992. Event-related potential indicators of the dynamic unconscious. *Consciousness and Cognition* 1:340-66.

Shevrin, H., Bond, J., Brakel, L., Hertel, R. & Williams, W. 1996. *Conscious and Unconscious Processes: Psychodynamic, Cognitive, and Neurophysiological Convergences*. Guilford Press.

Shevrin, H. 1998. The Freud-Rapaport theory of consciousness. In (R. Bornstein & J. Masling, eds) *Empirical Perspectives on the Psychoanalytic Unconscious*. American Psychological Association.

Slipp, S. 2000. Subliminal stimulation research and its implications for psychoanalytic theory and treatment. *Journal of the American Academy of Psychoanalysis* 28:305-320.

Smith, D. 2000. Freudian science of consciousness: Then and now. *Neuro-psychoanalysis* 2:38-45.

Solomon, R.C. 1974. Freud and "unconscious motivation". *Journal for the Theory of Social Behaviour* 4:191-216.

Solms, M. 1997. What is consciousness? *Journal of the American Psychoanalytic Association* 45:681-703.

Spence, D.P. & Holland, B. 1962. The restricting effects of awareness: A paradox and an explanation. *Journal of Abnormal and Social Psychology* 64:163-74.

Stein, D.J. (ed) 1997. *Cognitive Science and the Unconscious*. American Psychiatric Press.

Strauss, A. 1955. Unconscious mental processes and the psychosomatic concept.

International Journal of Psychoanalysis 36:307-19.

van der Waals, E.G. 1949. The psycho-analytical and the phenomenological concept of consciousness. International Journal of Psychoanalysis 30:207.

Wakefield, J.C. 1990. Why instinctual impulses can't be unconscious: An exploration of Freud's cognitivism. Psychoanalysis and Contemporary Thought 13:265-88.

Wakefield, J.C. 1991. Why emotions can't be unconscious: An exploration of Freud's essentialism. Psychoanalysis and Contemporary Thought 14:29-67.

Weinberger, J.& Weiss, J. 1997. Psychoanalytic and cognitive conceptions of the unconscious. In (D. Stein, ed) Cognitive Science and the Unconscious. American Psychiatric Press.

Westen, D. 1992. The cognitive self and the psychoanalytic self: Can we put our selves together? Psychological Inquiry 3:1-13.

Woody, J.M. & Phillips, J. 1995. Freud's project for a scientific psychology after 100 years: The unconscious mind in the era of cognitive neuroscience. Philosophy, Psychiatry, and Psychology 2:123-34.

Zilboorg, G. 1951. Variations in the scope of awareness. In (H. Abramson, ed) Problems of Consciousness: Transactions of the Second Conference. Josiah Macy Foundation.

6.2n Consciousness and Time

Allport, D.A. 1968. Phenomenal similarity and the perceptual moment hypothesis. British Journal of Psychology 59:395-406.

Banks, R. & Cappon, D. 1962. Effect of reduced sensory input on time perception. Perceptual and Motor Skills 14:74.

Block, R.A. 1979. Time and consciousness. In (G. Underwood & R. Stevens, eds) Aspects of Consciousness, Volume 1. Academic Press.

Block, R.A. (ed) 1990. Cognitive Models of Psychological Time. Lawrence Erlbaum.

Block, R.A. 1996. Psychological time and memory systems of the brain. In (J. Fraser & M. Soulsby, eds) Dimensions of Time and Life: The Study of Time, volume 8. International Universities Press.

Brown, J.W. 1990. Psychology of time awareness. Brain and Cognition 14:144-64.

Brown, J.W. 1991. Self and Process: Brain States and the Conscious Present. Springer-Verlag.

Brown, J. 2000. Mind and Nature: Essays on Time and Subjectivity. Whurr Publishers.

Cohen, J. 1954. The experience of time. Acta Psychologica 10:207-19.

Dennett, D.C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. Behavioral and Brain Sciences 15:183-201.

Eisler, H. 1975. Subjective duration and psychophysics. Psychological Review 82:429-50.

Fraser, J.T. (ed) 1989. Time and Mind: Interdisciplinary Issues.

International Universities Press.

Gooddy, W. 1967. Introduction to problems of time awareness. *Studium Generale* 20:33-41.

Hicks, R.E., Miller, G.W., Gaes, G., & Bierman, K. 1977. Concurrent processing demands and the experience of time-in-passing. *American Journal of Psychology* 90:431-46.

Hoagland, H. 1950. Consciousness and the chemistry of time. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the First Conference_*. Josiah Macy Foundation.

Hoagland, H. 1943. The chemistry of time. *Scientific Monthly* 56:56-61.

Knight, R.T. & Grabowecky, M. 1995. Escape from linear time: Prefrontal cortex and conscious experience. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.

Melges, F.T. 1989. Disorders of time and the brain in severe mental illness. In (J. Fraser, ed) *_Time and Mind: Interdisciplinary Issues_*. International Universities Press.

Michon, J.A. 1975. Time experience and memory processes. In (J. Fraser & N. Lawrence, eds) *_The Study of Time II_*. Springer-Verlag.

Michon, J.A. 1972. Processing of temporal information and the cognitive theory of time experience. In (J. Fraser, F. Haber, & G. Muller, eds) *_The Study of Time_*. Springer-Verlag.

Moiseeva, N.I. 19xx. Perception of time by human consciousness. *Chronobiologia* 15:301-317.

Natsoulas, T. 1993. The stream of consciousness: II. William James's specious present. *Imagination, Cognition and Personality* 12:367-385.

Newman, M.A. 1982. Time as an index of expanding consciousness with age. *Nursing Research* 31:290-293.

Orme, J.E. 1969. *_Time, Experience and Behaviour_*. Illife.

Ornstein, R.E. 1969. *_On the Experience of Time_*. Harmondsworth.

Poppel, E. 1988. *_Mindworks: Time and Conscious Experience_*. Harcourt Brace Jovanovich.

Poppel, E. & Schwender, D. 1993. Temporal mechanisms of consciousness. *International Anesthesiology Clinics* 31:27-38.

Proust, J. 1994. Time and conscious experience. In (C.C. Gould, ed) *_Artifacts, Representations, and Social Practice_*. Kluwer.

Reidhead, V.A. & Wolford, J.B. 1998. Context, conditioning, and meaning of time-consciousness in a Trappist monastery. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Robertson, T.B. 1923. Consciousness and the sense of time. *Scientific Monthly* 16:649-657.

Sanders, S.A. 1986. Development of a tool to measure subjective time experience. *Nursing Research* 35:178-182.

Schaltenbrand, G. 1967. Consciousness and time. *Annals of the New York Academy of Sciences* 138:632-645.

Strong, C.A. 1896. Consciousness and time. *Psychological Review* 3:149-57.

Stroud, J.M. 1967. The fine structure of psychological time. *Annals of the New York Academy of Sciences* 138:623-631.

Stroud, J.M. 1957. The fine structure of psychological time. In (H. Quastler, ed) *_Information Theory in Psychology: Problems and Methods_*. Free Press.

6.2o Self-Consciousness

Asendorpf, J.B., Warkentin, V., & Baudonniere, P. 1996. Self-awareness and other-awareness. II: Mirror self-recognition, social contingency awareness, and synchronic imitation. *Developmental Psychology* 32q:313-321.

Butterworth, G. 1995. The self as an object of consciousness in infancy. In (P. Rochat, ed) *_The Self in Infancy: Theory and Research_*. Elsevier.

Carver, C. & Scheier, M.F. 1983. Self-awareness and the self-regulation of behaviour. In (G. Underwood, ed) *_Aspects of Consciousness, Volume 3: Awareness and Self-Awareness_*. Academic Press.

Cheeks, J.M. & Briggs, S.R. 1982. Self-consciousness and aspects of personality. *Journal of Research in Personality* 16:401-8.

Cooney, B. 1979. The neural basis of self-consciousness. *Nature and System* 1:16-31.

Duval, S. & Wicklund, R.A. 1972. *_A Theory of Objective Self-Awareness_*. Academic Press.

Dymond, S. & Barnes, D. 1997. Behavior-analytic approaches to self-awareness. *Psychological Record* 47:181-200.

Feinberg, T.E. 1997. Some interesting perturbations of the self in neurology. *Seminars in Neurology* 17:129-35.

Fenigstein, A., Scheier, M.F. & Buss. A.H. 1975. Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology* 43:522-27.

Fenigstein, A. 1997. Self-consciousness and its relation to psychological mindedness. In (M. McCallum & W. Piper, eds) *_Psychological Mindedness: A Contemporary Understanding_*. Lawrence Erlbaum.

Frith, C.D. 1996. The role of the prefrontal cortex in self-consciousness: the case of auditory hallucinations. *Philosophical Transactions of the Royal Society of London* B351:1505-12.

Freeman, W. & Watts, J.W. 1941. The frontal lobes and consciousness of self. *Psychosomatic Medicine* 3:111-19.

Gallup, G.G. 1998. Self-awareness and the evolution of social intelligence. *Behavioural Processes* 42:239-247.

Gardiner, J. 2000. On the objectivity of subjective experiences and autonoetic and noetic consciousness. In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.

Hart, D. & Fegley, S. 1997. The development of self-awareness and self-understanding in cultural context. In (U. Neisser & D. Jopling, eds) *_The Conceptual Self in Context_*. Cambridge University Press.

- Hart, D. & Fegley, S. 1994. Social imitation and the emergence of a mental model of self. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- James, W. 1890. The consciousness of self. In *_The Principles of Psychology_*.
- Jaynes, J. 1976. *_The Origins of Consciousness in the Breakdown of the Bicameral Mind_*. Houghton Mifflin.
- Kessel, F.S, Cole, P.M. & Johnson, D.L. (eds) *_Self and Consciousness: Multiple Perspectives_*. Lawrence Erlbaum.
- Kihlstrom, J.F. 1997. Consciousness and me-ness. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.
- Kihlstrom, J.F. & Klein, S.B. 1997. Self-knowledge and self-awareness. In (J. Snodgrass, R. Thompson, eds) *_The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept_*. New York Academy of Sciences.
- Kinsbourne, M. 1995. Awareness of one's own body: An attentional theory of its nature, development, and brain basis. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *_The Body and the Self_*. MIT Press.
- Kinsbourne, M. 1998. Representations in consciousness and the neuropsychology of insight. In (X. Amador & A. David, eds) *_Insight and Psychosis_*. Oxford University Press.
- Kunzendorf, R.G. 1988. Self-consciousness as the monitoring of cognitive states: A theoretical perspective. *Imagination, Cognition and Personality* 7:3-22.
- Kunzendorf, R.G., Beltz, S.M. & Tymowicz, G. 1992. Self-awareness in autistic subjects and deeply hypnotized subjects: Dissociation of self-concept versus self-consciousness. *Imagination, Cognition and Personality* 11:129-41.
- Kunzendorf, R. 2000. Individual differences in self-conscious source monitoring: Theoretical, experimental, and clinical considerations. In (R. Kunzendorf & B. Wallace, eds) *_Individual Differences in Conscious Experience_*. John Benjamins.
- Levine, B. 2000. Self-regulation and autonoetic consciousness. In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.
- Lewis, M. 1991. Ways of knowing: Objective self-awareness or consciousness. *Developmental Review* 11:231-43.
- Lewis, M. 1994. Myself and me. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- Markova, I. 1990. The development of self-consciousness: Baldwin, Mead, and Vygotsky. In (J. Faulconer & R. Williams, eds) *_Reconsidering Psychology_*. Duquesne University Press.
- Mitchell, R.W. 1993. Mental models of mirror self-recognition: Two theories. *New Ideas in Psychology* 11:295-325.
- Mitchell, R.W. 1994. Multiplicities of self. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.

Mitchell, R.W. 1997. A comparison of the self-awareness and kinesthetic-visual matching theories of self-recognition: Autistic children and others. In (J. Snodgrass, R. Thompson, eds) *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.

Mollon, P. 1987. Self-awareness, self-consciousness, and preoccupation with self. In (K. Yardley & T. Honess, eds) *Self and Identity: Psychosocial Perspectives*. Wiley.

Morin, A. & Everett, J. 1990. Inner speech as a mediator of self-awareness, self-consciousness, and self-knowledge: An hypothesis. *New Ideas in Psychology* 8:337-56.

Nasby, W. 1989. Private self-consciousness, self-awareness, and the reliability of self-reports. *Journal of Personality and Social Psychology* 56:950-7.

Neisser, U. 1991. Five kinds of self-knowledge. *Philosophical Psychology* 1:35-59.

Neisser, U. 1992. The development of consciousness and the acquisition of self. In (F. Kessel, P. Cole, & D.L. Johnson, eds) *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.

Parker, S.T., Mitchell, R.M., & Boccia, M.L. 1994. *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Sass, L. 2000. Schizophrenia, self-experience, and the so-called "negative symptoms": Reflections on hyperreflexivity. In (D. Zahavi, ed) *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience*. John Benjamins.

Shotter, J. 1983. Consciousness and self-consciousness: Inner games and alternative realities. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Shrauger, J.S. & Osberg, T.M. 1983. Self-awareness: The ability to predict one's subsequent behaviour. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Siegrist, M. 1995. Inner speech as a cognitive process mediating self-consciousness and inhibiting self-deception. *Psychological Reports* 76:259-65.

Snodgrass, J.G. & Thompson, R.L. (eds) 1997. *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.

Titchener, E.B. 1911. A note on the consciousness of self. *American Journal of Psychology* 22:540-52.

Watson, J.S. 1994. Detection of self: The perfect algorithm. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Watson, P.J., Morris, R.J., Ramsey, A. Hickman, S.E. 1996. Further contrasts between self-reflectiveness and internal state awareness factors of private self-consciousness. *Journal of Psychology* 130:183-92.

6.2p Development of Consciousness -----

Anderson, J.R. 1984. The development of self-recognition: A review.

Developmental Psychobiology 17:35-49.

Brainerd, C.J., Stein, L.M., & Reyna, V.F. 1998. On the development of conscious and unconscious memory. *Developmental Psychology* 34:342-357.

Briskin, A.S. 1974. A developmental model of self-awareness. *Counseling and Values* 18:79-85.

Burgess, J.A. & Tawia, S.A. 1996. When did you first begin to feel it? Locating the beginnings of human consciousness? *Bioethics* 10:1-26.

Butterworth, G. 1995. The self as an object of consciousness in infancy. In (P. Rochat, ed) *The Self in Infancy: Theory and Research*. Elsevier.

Davis, L.H. 1989. Self-consciousness in chimps and pigeons. *Philosophical Psychology* 2:249-59.

Flavell, J.H. 1993. Young children's understanding of thinking and consciousness. *Current Directions in Psychological Science* 2:40-43.

Flavell, J.H., Green, F.L., & Flavell, E.R. 1993. Children's understanding of the stream of consciousness. *Child Development* 64:387-398.

Flavell, J.H., Green, F.L., Flavell, E.R. & Grossman, J.B. 1997. The development of children's knowledge about inner speech. *Child Development* 68:39-47.

Flavell, J.H., Green, F.L., Flavell, E.R. 1995. The development of children's knowledge about attentional focus. *Developmental Psychology* 31:706-12.

Flavell, J., Green, F., & Flavell, E. 2000. Development of children's awareness of their own thoughts. *Journal of Cognition & Development* 1:97-112.

Foulkes, D. 1999. *Children's Dreaming and the Development of Consciousness*. Harvard University Press.

Gallagher, S. & Meltzoff, A. 1996. The earliest sense of self and others: Merleau-Ponty and recent developmental studies. *Philosophical Psychology* 9:211-33.

Gopnik, A. & Meltzoff, A.N. 1994. Minds, bodies, and persons: Young children's understanding of the self and others as reflected in imitation and theory of mind research. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Griffin, S. 1991. Young children's awareness of their inner world: A neo-structural analysis of the development of intrapersonal intelligence. In (R. Case, ed) *The Mind's Staircase: Exploring the Conceptual Underpinnings of Children's Thought and Knowledge*. Lawrence Erlbaum.

Kagan, J. 1981. *The Second Year: The Emergence of Self-Awareness*. Harvard University Press.

Kuhn, D. 2000. Metacognitive development. *Current Directions in Psychological Science* 9:178-181.

Lewis, M. 1990. The development of intentionality and the role of consciousness. *Psychological Inquiry* 1:231-247.

Lewis, M. 1991. Ways of knowing: Objective self-awareness or consciousness. *Developmental Review* 11:231-43.

Lunzer, E.A. 1979. The development of consciousness. In (G. Underwood &

R. Stevens, eds) *_Aspects of Consciousness_*. Academic Press.

Marbach, E. 1987. Laws of consciousness as norms of mental development. In (B. Inhelder, D. de Caprona, & A. Cornu-Wells, eds) *_Piaget Today_*. Lawrence Erlbaum.

Markova, I. 1990. The development of self-consciousness: Baldwin, Mead, and Vygotsky. In (J. Faulconer & R. Williams, eds) *_Reconsidering Psychology_*. Duquesne University Press.

McCune, L. 1993. The development of play as the development of consciousness. In (M. Bornstein & A. O'Reilly, eds) *_The Role of Play in the Development of Thought_*. Jossey-Bass.

Mounoud, P. 1990. Consciousness as a necessary transitional phenomenon in cognitive development. *Psychological Inquiry* 1:253-58.

Neisser, U. 1992. The development of consciousness and the acquisition of skill. In (F. Kessel, P. Cole, & D. Johnson, eds) *_Self and Consciousness: Multiple Perspectives_*. Lawrence Erlbaum.

Parker, S.T., Mitchell, R.M., & Boccia, M.L. 1994. *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.

Piaget, J. 1954. The problem of consciousness in child psychology: Developmental changes in awareness. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the Fourth Conference_*. Josiah Macy Foundation.

van Eenwyk, J.R. 1996. Chaotic dynamics and the development of consciousness. In (E. MacCormac & M. Stamenov, eds) *_Fractals of Brain, Fractals of Mind: In Search of a Symmetry Bond_*. John Benjamins.

Wheeler, M. 2000. Varieties of consciousness and memory in the developing child. In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.

Wilber, K. 1979. A developmental view of consciousness. *Journal of Transpersonal Psychology* 11:1-21.

Zelazo, P. D. 1996. Towards a characterization of minimal consciousness. *New Ideas in Psychology* 14:63-80.

Zelazo, P.R. & Zelazo, P.D. 1998. The emergence of consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Zelazo, P. 2000. Self-reflection and the development of consciously controlled processing. In (P. Mitchell & K. Riggs, eds) *_Children's Reasoning and the Mind_*. Psychology Press/Taylor & Francis.

6.2q The Stream of Consciousness

Antrobus, J.S., Singer, J.L., & Greenberg, S. 1966. Studies in the stream of consciousness: Experimental enhancement and suppression of spontaneous cognitive processes. *Perceptual and Motor Skills* 23:399-417.

Baars, B.J. 1993. How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity? In *_Experimental and Theoretical Studies of Consciousness_* (Ciba Foundation Symposium 174).

- Bakan, P. 1978. Two streams of consciousness: A typological approach. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.
- Bonanno, G.A. & Singer, J.L. 1993. Controlling one's stream of thought through perceptual and reflective processing. In (D. Wegner & J. Pennebaker, eds) *Handbook of Mental Control*. Prentice-Hall.
- Capek, M. 1950. Stream of consciousness and "duree reelle." *Philosophy and Phenomenological Research* 10:331-353.
- Diaz, J. 1996. The stream revisited: A process model of phenomenological consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.
- Dainton, B. 2000. *Stream of Consciousness: Unity and continuity in conscious experience*. Routledge.
- Flanagan, O. 1992. The stream of consciousness. In *Consciousness Reconsidered*. MIT Press.
- Flavell, J.H., Green, F.L., & Flavell, E.R. 1993. Children's understanding of the stream of consciousness. *Child Development* 64:387-398.
- Gurwitsch, A. 1943. William James' theory of the "transitive parts" of the stream of consciousness. *Philosophy and Phenomenological Research* 3:449-477.
- James, W. 1990. The stream of thought. In *Principles of Psychology*.
- James, W. 1992. The stream of consciousness. In *Psychology: The Briefer Course*.
- Klinger, E. 1978. Modes of normal conscious flow. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.
- Mueller, E.T. 1990. *Daydreaming in Humans and Machines: A Computer Model of the Stream of Thought*. Ablex.
- Natsoulas, T. 1987. The six basic concepts of consciousness and William James' stream of thought. *Imagination, Cognition, and Personality* 6:289-319.
- Natsoulas, T. 1988. Sympathy, empathy, and the stream of consciousness. *Journal for the Theory of Social Behaviour* 18:169-195.
- Natsoulas, T. 1992-1996. The stream of consciousness: Parts I-XVI. *Imagination, Cognition, and Personality* 12:3-21, 12:367-85, 13:73-90, 13:229-47, 13:347-66, 14:59-77, 14:131-49, 14:333-52, 15:171-91, 16:161-80, 16:281-300, 17:45-64, 17:123-40, 17:229-47.
- Penfield, W. 1955. The permanent record of the stream of consciousness. *Acta Psychologica* 11:47-69.
- Petchkovsky, L. 2000. 'Stream of consciousness' and 'ownership of thought' in indigenous people in Central Australia. *Journal of Analytical Psychology* 45:577-597.
- Pollio, H.R. 1990. The stream of consciousness since James. In (M. Johnson & T. Henley, eds) *Reflections on "The Principles of Psychology": William James after a Century*. Lawrence Erlbaum.
- Pope, K.S. & Singer, J.L. (eds) 1978. *The Stream of Consciousness: Scientific Investigations into the Flow of Human Experience*. Plenum Press.

Pope, K.S. 1978. How gender, solitude, and posture influence the stream of consciousness. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Rychlak, J.F. 1978. The stream of consciousness: Implications for a humanistic psychological theory. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Schuetz, A. 1940. William James's concept of the stream of thought, phenomenologically interpreted. *Journal of Philosophy* 37:673-74.

Singer, J.L. 1974. Daydreaming and the stream of thought. *American Scientist* 62:417-425.

Singer, J.L. 1975. Navigating the stream of consciousness: Research in daydreaming and related inner experience. *American Psychologist* 30:727-738.

Singer, J.L. 1978. Experimental studies of daydreaming and the stream of thought. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Singer, J.L. 1998. Daydreams, the stream of consciousness, and self-representations. In (R. Bornstein & J. Masling, eds) *Empirical Perspectives on the Psychoanalytic Unconscious*. American Psychological Association.

Strange, J.R. 1978. A search for the sources of the stream of consciousness. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

6.2r Foundational Issues

Baars, B.J. 1986. What is a theory of consciousness a theory of? The search for criterial constraints on theory. *Imagination, Cognition, and Personality* 1:3-24.

Baars, B.J. 1994. A thoroughly empirical approach to consciousness. *Psyche* 1.

Baars, B.J. 1996. Understanding subjectivity: Global workspace theory and the resurrection of the observing self. *Journal of Consciousness Studies* 3:211-17.

Benoit, P.J. & Benoit, W.L. 1986. Consciousness: The mindlessness/mindfulness and verbal report controversies. *Western Journal of Speech Communication* 50:41-63.

Bindra, D. 1970. The problem of subjective experience. *Psychological Review* 77:581-84.

Blanshard, B. & Skinner, B.F. 1966. The problem of consciousness: A debate. *Philosophy and Phenomenological Research* 27:317-37.

Carlson, R.A. 1992. Starting with consciousness. *American Journal of Psychology* 105:598-604.

Casler L. 1976. The "consciousness problem" is not the problem. *Perceptual and Motor Skills* 42:227-32.

Hebb, D.O. 1954. The problem of consciousness and introspection. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.

Kihlstrom, J.F. 1987. What this discipline needs is a good ten-cent taxonomy of consciousness. *Canadian Psychology* 28:116-118.

- Merikle, P.M. 1984. Toward a definition of awareness. *Bulletin of the Psychonomic Society* 22:449-50.
- Natsoulas, T. 1974. The subjective, experiential element in perception. *Psychological Bulletin* 81:611-31.
- Natsoulas, T. 1981. Basic problems of consciousness. *Journal of Personality and Social Psychology* 41:132-78.
- Natsoulas, T. 1990. Is consciousness what psychologists actually examine? *American Journal of Psychology* 105:363-84.
- Revonsuo, A. 1993. Is there a ghost in the cognitive machinery? *Philosophical Psychology* 6:387-405.
- Rychlak, J.F. 1997. *_In Defense of Human Consciousness_*. American Psychological Association.
- Tolman, E.C. 1935. Psychology versus immediate experience. *Philosophy of Science* 2:356-80.
- Wilson, D.L. 1978. Brain mechanisms, consciousness, and introspection. In (A. Sugarman & R. Tarter, eds) *_Expanding Dimensions of Consciousness_*. Springer.
- Zener, K. 1952. Significance of the experience of the individual for the science of psychology. *Minnesota Studies in the Philosophy of Science* 2:354-69.

6.2s Consciousness and Psychology, Misc -----

- Aurell, G. 1979. Perception: A model comprising two modes of consciousness. *Perceptual and Motor Skills* 49:431-44.
- Aurell, G. 1989. Man's triune conscious mind, parts I, II, and III. *Perceptual and Motor Skills* 68:747-54, 78:31-39, 81:463-66
- Bolton, N. 1983. Forms of awareness. In (G. Underwood, ed) *_Aspects of Consciousness, Volume 3: Awareness and Self-Awareness_*. Academic Press.
- Bowden, E.M. 1997. The effect of reportable and unreportable hints on anagram solution and the aha! experience. *Consciousness and Cognition* 6:545-573.
- Davidson, J.M. & Davidson, R.J. (eds) 1980. *_The Psychobiology of Consciousness_*. Plenum.
- Davidson, R., Schwartz, G. & Shapiro, D. (eds) 1983. *_Consciousness and Self-Regulation_*. Plenum.
- Delacour, J. 1997. Object perception and recognition: A model for the scientific study of consciousness. *Theory and Psychology* 7:257-62.
- Dulany, D.E. 1991. Conscious representation and thought systems. In (R. Wyer & T. Srull, eds) *_The Content, Structure, and Operation of Thought Systems_*. Lawrence Erlbaum.
- Dulany, D.E. 1997. Consciousness in the explicit (deliberative) and implicit (evocative). In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.
- Farthing, G.W. 1992. *_The Psychology of Consciousness_*. Prentice Hall.

- Frith C.D. 1979. Consciousness, information processing and schizophrenia. *British Journal of Psychiatry* 134:225-35.
- Greenberg, G. & Tobach, E. (eds) 1987. Cognition, language, and consciousness: Integrative levels. Lawrence Erlbaum.
- Higgins, E.E. & Bargh, J.A. 1992. Unconscious sources of subjectivity and suffering: Is consciousness the solution? In (L. Martin & A. Tesser, eds) *The Construction of Social Judgments*. Lawrence Erlbaum
- Hilgard, E.R. 1977. *Divided Consciousness: Multiple Controls in Human Thought and Action*. Wiley.
- Hilgard, E.R. 1977. Controversies over consciousness and the rise of cognitive psychology. *Australian Psychologist* 12:7-27.
- Hilgard E.R. 1977. The problem of divided consciousness: A neodissociation interpretation. *Annals of the New York Academy of Sciences* 296:48-59.
- Hilgard, E.R. 1980. Consciousness in contemporary psychology. *Annual Review of Psychology* 31:1-26.
- Hilgard, E.R. 1992. Divided consciousness and dissociation. *Consciousness and Cognition* 1:16-31.
- Hirst, W. 1995. Cognitive aspects of consciousness. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Humphrey, N. 1992. *A History of the Mind*. Simon and Schuster.
- Jonassen, D.H. 1979. Video-mediated, objective self-awareness, self-perception, and locus of control. *Perceptual and Motor Skills* 48:255-265.
- Kihlstrom, J.F. 1993. The continuum of consciousness. *Consciousness and Cognition* 2:334-54.
- Lashley, K.S. 1923. The behavioristic interpretation of consciousness. *Psychological Review* 30:237-72.
- Lewicki, P., Czyzewska, M. & Hill, T. 1997. Cognitive mechanisms for acquiring "experience": The dissociation between conscious and nonconscious cognition. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Mandler, G. 1975. Consciousness: respectable, useful, and probably necessary. In (R. Solso, ed) *Information Processing and Cognition*. Lawrence Erlbaum.
- Mandler, G. 1988. Problems and direction in the study of consciousness. In (M. Horowitz, ed) *Psychodynamics and Cognition*. University of Chicago Press.
- Mandler, G. 1992. Toward a theory of consciousness. In (H.G. Geissler, S.W. Link, & J.T. Townsend, eds) *Cognition, Information Processing, and Psychophysics: Basic Issues*. Lawrence Erlbaum.
- Mandler, G. 1997. Consciousness redux. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Mandler, G. & Nakamura, Y. 1987. Aspects of consciousness. *Personality and Social Psychology Bulletin* 13:299-313.
- Natsoulas, T. 1984. Personality and consciousness: A theoretical essay. *Cognition and Brain Theory* 7:135-66.
- Oakley, D.A. & Eames, L.C. 1986. The plurality of consciousness. In

(D. Oakley, ed) *Mind and Brain*. Methuen.

Oatley, K. 1988. On changing one's mind: A possible function of consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Ornstein, R.E. (ed) 1974. *The Nature of Human Consciousness: A Book of Readings*. Viking Press.

Ornstein, R.E. 1977. *The Psychology of Consciousness*. Harcourt Brace Jovanovich.

Posner, M.I. & Klein, M. 1973. On the functions of consciousness. In (S. Kornblum, ed) *Attention and Performance*, vol 4. Academic Press.

Schwartz, G. & Shapiro, D. (eds) 1976. *Consciousness and Self-regulation*. Plenum.

Schwartz, G. 2000. Individual differences in subtle awareness and levels of awareness: Olfaction as a model system. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Schwarz, N. & Clore, G.L. 1996. Feelings and phenomenal experiences. In *Social Psychology: Handbook of Basic Principles*. Guilford Press.

Shallice, T. 1991. The revival of consciousness in cognitive science. In (W. Kessen, A. Ortony, & F. Craik, eds) *Memories, Thoughts, and Emotions: Essays in Honor of George Mandler*. Lawrence Erlbaum.

Solso, R. (ed) 1975. *Information Processing and Consciousness*. Lawrence Erlbaum.

Sperry, R.W. 1987. Structure and significance of the consciousness revolution. *Journal of Mind & Behavior* 8:37-65.

Sperry, R.W. 1995. The riddle of consciousness and the changing scientific worldview. *Journal of Humanistic Psychology* 35:7-33.

Stout, M. 2001. *The Myth of Sanity: Divided Consciousness and the Promise of Awareness*. Viking/Penguin Books.

Underwood, G. & Stevens, R. (eds) 1979. *Aspects of Consciousness: Volume 1, Psychological Issues*. Academic Press.

Underwood, G. & Stevens, R. (eds) 1981. *Aspects of Consciousness: Volume 2, Structural Issues*. Academic Press.

Underwood, G. & Stevens, R. (eds) 1982. *Aspects of Consciousness: Volume 3, Awareness and Self-Awareness*. Academic Press.

Underwood, G. & Stevens, R. (eds) 1984. *Aspects of Consciousness: Volume 4, Clinical Issues*. Academic Press.

Underwood, G. (ed) 1996. *Implicit Cognition*. Oxford University Press.

von der Malsburg, C. 1986. Am I thinking assemblies? In (G. Palm & A. Aertsen, eds) *Brain Theory*. Springer.

von der Malsburg, C. 1997. The coherence definition of consciousness. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.

Wallace, B. & Fisher, L. 2000. Biological rhythms and individual differences in consciousness. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in*

Conscious Experience_. John Benjamins.

Wilks, Y. 1984. Machines and consciousness. In (C. Hookway, ed) Minds, Machines and Evolution. Cambridge University Press.

6.3 Consciousness and Physics

6.3a The Interpretation of Quantum Mechanics

Albert, D. & Loewer, A. 1988. Interpreting the many-worlds interpretation. *Synthese* 77:195-213.

Butterfield, J. 1996. Whither the minds? *British Journal for the Philosophy of Science* 47:200-??.

Butterfield, J. 1998. Quantum curiosities of psychophysics. In (J. Cornwell, ed) Consciousness and Human Identity. Oxford University Press.

Byrne, A. & Hall, N. 1999. Chalmers on consciousness and quantum mechanics. *Philosophy of Science* 66:370-90.

Goertzel, B. 1992. Quantum theory and consciousness. *Journal of Mind and Behavior* 13:29-36.

Goswami, A. 1989. The idealistic interpretation of quantum mechanics. *Physics Essays* 2:385-400.

Goswami, A. 1990. Consciousness in quantum physics and the mind-body problem. *Journal of Mind and Behavior* 11:75-96.

Klein, S. 1991. The duality of psycho-physics. In (A. Gorea, ed) Representations of Vision. Cambridge University Press.

Lehner, C. 1997. What it feels like to be in a superposition, and why: Consciousness and the interpretation of Everett's quantum mechanics. *Synthese* 110:191-216.

Lockwood, M. 1989. Mind, Brain, and the Quantum. Oxford University Press.

Lockwood, M. 1996. Many-minds interpretations of quantum mechanics. *British Journal for the Philosophy of Science* 47:159-88.

Mulhauser, G. 1995. Materialism and the "problem" of quantum measurement. *Minds and Machines* 5:207-17.

Mulhauser, G. 1995. On the end of a quantum-mechanical romance. *Psyche* 2(19).

Page, D.N. 1995. Attaching theories of consciousness to Bohmian quantum mechanics. Manuscript.

Page, D.N. 1996. Sensible quantum mechanics: Are probabilities only in the mind? *International Journal of Modern Physics* D5:583-96.

Penrose, R. 1987. Quantum physics and conscious thought. In (B. Hiley & D. Peat, eds) Quantum Implications: Essays in Honour of David Bohm. Methuen.

Shanks, N. 1995. Minds, brains, and quantum mechanics. *Southern Journal of Philosophy* 33:243-60.

Squires, E.J. 1991. One mind or many? A note on the Everett interpretation of quantum theory. *Synthese* 89:283-6.

Squires, E.J. 1993. Quantum theory and the relation between the conscious mind and the physical world. *Synthese* 97:109-23.

Squires, E.J. 1994. Quantum theory and the need for consciousness. *Journal of Consciousness Studies* 1:201-4.

Squires, E.J. 1998. Why are quantum theorists interested in consciousness? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Stapp, H.P. 1991. Quantum propensities and the brain-mind connection. *Foundations of Physics* 21:1451-77.

Stapp, H.P. 1993. *Mind, Matter, and Quantum Mechanics*. Springer-Verlag.

Stapp, H.P. 1995. Why classical mechanics cannot accommodate consciousness but quantum mechanics can. *Psyche* 2(5).

Stapp, H.P. 1998. The evolution of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Wigner, E. 1961. Remarks on the mind-body problem. In (I. Good, ed) *The Scientist Speculates*. Heineman.

Woo, C.H. 1981. Consciousness and quantum interference: An experimental approach. *Foundations of Physics* 11:933-44.

6.3b Quantum Mechanisms of Consciousness

Bass, L. 1975. A quantum-mechanical mind-body interaction. *Foundations of Physics* 5:159-72.

Beck, F. & Eccles, J. 1992. Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Science USA* 89:11357-61.

Beck, F. 1998. Synaptic transmission, quantum-state selection, and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Berezin, A.A. 1992. Correlated isotopic tunneling as a possible model for consciousness. *Journal of Theoretical Biology* 154:415-20.

Dyer, M.G. 1994. Quantum physics and consciousness, creativity, computers: A commentary on Goswami's quantum-based theory of consciousness and free will. *Journal of Mind and Behavior* 15:265-90.

Eccles, J.C. 1986. Do mental events cause neural events analogously to the probability fields of quantum mechanics? *Proceedings of the Royal Society of London B* 227:411-28.

Germine, M. 1991. Consciousness and synchronicity. *Medical Hypotheses* 36:277-83.

Globus, G. 1997. Nonlinear brain systems with nonlocal degrees of freedom. *Journal of Mind and Behavior*.

Globus, G. 1998. Self, cognition, qualia, and world in quantum brain dynamics. *Journal of Consciousness Studies* 5:34-52.

Grush, R. & Churchland, P. 1995. Gaps in Penrose's toiling. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Hameroff, S.R. 1994. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1:91-118.

Hameroff, S.R. & Penrose, R. 1996. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Hameroff, S.R. & Penrose, R. 1996. Conscious events as orchestrated space-time selections. *Journal of Consciousness Studies* 3:36-53. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.

Hameroff, S.R. & Scott, A. 1998. A Sonoran afternoon: A dialogue on quantum mechanics and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Jibu, M. & Yasue, K. 1995. *Quantum Brain Dynamics and Consciousness: An Introduction*. John Benjamins.

Jibu, M. & Yasue, K. 1997. Magic without magic: Meaning of quantum brain dynamics. *Journal of Mind and Behavior*.

King, C. 1997. Chaos, quantum mechanics, and the conscious brain. *Journal of Mind and Behavior*.

Lahav, R. & Shanks, N. 1992. How to be a scientifically respectable 'property dualist'. *Journal of Mind and Behavior* 13:211-32.

Marshall, I.N. 1989. Consciousness and Bose-Einstein condensates. *New Ideas in Psychology* 7:73-83.

Marshall, I.N. 1995. Some phenomenological implications of a quantum model of consciousness. *Minds and Machines* 5:609-20.

Penrose, R. 1994. Mechanisms, microtubules, and the mind. *Journal of Consciousness Studies* 1:241-49.

Scott, A. 1996. On quantum theories of the mind. *Journal of Consciousness Studies* 3:484-91.

Stapp, H.P. 1985. Consciousness and values in the quantum universe. *Foundations of Physics* 15:35-47.

Stapp, H.P. 1994. Theoretical model of a purported empirical violation of the predictions of quantum mechanics. *Physical Review A* 50:18-22.

Stapp, H.P. 1995. The hard problem: A quantum approach. *Journal of Consciousness Studies* 3:194-210. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.

Stapp, H.P. 1997. Science of consciousness and the hard problem. *Journal of Mind and Behavior* 18:171-93.

Triffet, T. & Green, H.S. 1996. Consciousness: Computing the uncomputable. *Mathematical and Computational Modelling* 24:37-56.

Wolf, F.A. 1996. On the quantum mechanics of dreams and the emergence of self-awareness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Zohar, D. 1995. A quantum-mechanical model of consciousness and the emergence of 'I'. *Minds and Machines* 5:597-607.

Zohar, D. 1996. Consciousness and Bose-Einstein condensates. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

6.3c Consciousness and Physics, Misc

Bilodeau, D. 1996. Physics, machines, and the hard problem. *Journal of Consciousness Studies* 3:386-401. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.

Bohm, D.J. 1986. A new theory of the relationship of mind and matter. *Journal of the American Society for Psychical Research* 80:113-35.

Burns, J. 1990. Contemporary models of consciousness, parts I & II. *Journal of Mind and Behavior* 11:153-171 & 12:407-420.

Clarke, C.J.S. 1995. The nonlocality of mind. *Journal of Consciousness Studies* 2:231-40. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.

Culbertson, J. 1982. *Consciousness: Natural and Artificial*. Libra.

de Silva, F. 1996. Consciousness and special relativity. *IEEE Engineering in Medicine and Biology Magazine* 15:21-26.

Dyer, M.G. 1994. Quantum physics and consciousness, creativity, and computers. *Journal of Mind and Behavior* 15:265-90.

Elitzur, A. 1996. Time and consciousness: The uneasy bearing of relativity on the mind-body problem. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Herbert, N. 1993. *Elemental Mind: Human Consciousness and the New Physics*. Dutton.

Ho, M.W. 1997. Quantum coherence and conscious experience. *Kybernetes* 26:265-76.

Hodgson, D. 1988. *The Mind Matters: Consciousness and Choice in a Quantum World*. Oxford: Oxford University Press.

Hodgson, D. 1996. Nonlocality, local indeterminism, and consciousness. *Ratio* 9:1-22.

Nair, R. 1991. Quantum physics and the philosophy of mind: An essay review. *Journal of Scientific and Industrial Research* 50:66975.

Nunn, C.M.H., Clarke, C.J.S. & Blott, B.H. 1994. Collapse of a quantum field may affect brain function. *Journal of Consciousness Studies* 1:127-39.

Nunn, C.M.H. 1996. On the geometry of consciousness. *Journal of Consciousness Studies* 3:477-83.

Penrose, R. 1989. *The Emperor's New Mind*. Oxford University Press.

Penrose, R. 1994. *Shadows of the Mind*. Oxford University Press.

Penrose, R. 1997. *The Large, the Small, and the Human Mind*. Cambridge University Press.

Squires, E. 1990. *Conscious Mind in the Physical World*. Adam Hilger.

Zohar, D. & Marshall, I. 1990. *The Quantum Self*. Morrow.

6.4 Consciousness and Science, Misc

6.4a Evolution of Consciousness

Arhem, P. & Liljenstrom, H. 1997. On the coevolution of consciousness and cognition. *Journal of Theoretical Biology* 187:601-12.

Barlow, H.B. 1980. Nature's joke: A conjecture on the biological role of consciousness. In (B. Josephson & V. Ramachandran, eds) *_Consciousness and the Physical World_*. Pergamon Press.

Barlow, H.B. 1987. The biological role of consciousness. In (C. Blakemore & S. Greenfield, eds) *_Mindwaves_*. Blackwell.

Cairns-Smith, A.G. 1996. *_Evolving the Mind: On the Nature of Matter and the Origin of Consciousness_*. Cambridge University Press.

Coan, R.W. 1989. Alternative views on the evolution of consciousness. *Journal of Human Psychology* 29:167-99.

Cotterill, R. 2000. Did consciousness evolve from self-paced probing of the environment, and not from reflexes? *Brain and Mind* 1:283-298.

Crook, J.H. 1980. *_The Evolution of Human Consciousness_*. Oxford University Press.

Donald, M. 1995. The neurobiology of human consciousness: An evolutionary approach. *Neuropsychologia* 33:1087-1102.

Eccles, J.C. 1992. Evolution of consciousness. *Proceedings of the National Academy of Sciences USA* 89:7320-24.

Calvin, W.H. 1991. *_The Ascent of Mind: Ice Age Climates and the Evolution of Intelligence_*. Bantam Books.

Crook, J.H. 1980. *_The Evolution of Human Consciousness_*. Oxford University Press.

Dennett, D.C. 1986. Julian Jaynes' software archaeology. *Canadian Psychology* 27:149-54.

Dewart, L. 1989. *_Evolution and Consciousness: The Role of Speech in the Origin and Development of Human Nature_*. University of Toronto Press.

Glynn, I.M. 1993. The evolution of consciousness: William James' unresolved problem. *Biological Reviews of the Cambridge Philosophical Society* 68:599-616.

Hameroff, S.R. 1998. Did consciousness cause the Cambrian evolutionary explosion? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Humphrey, N. 1992. *_A History of the Mind: Evolution and the Birth of Consciousness_*. Simon and Schuster.

Jaynes, J. 1976. *_The Origins of Consciousness in the Breakdown of the Bicameral Mind_*. Houghton Mifflin.

Jonker A. 1987. The origin of the human mind. A speculation on the emergence of language and human consciousness. *Acta Biotheoretica* 36:129-77.

King, J.E., Rumbaugh, D.M. & Savage-Rumbaugh, E.S. 1998. Evolution of intelligence, language, and other emergent processes for consciousness: A comparative perspective. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Lindahl, B.I.B. 1997. Consciousness and biological evolution. *Journal of Theoretical Biology* 187:613-29.

Nichols, S. & Grantham, T. 2000. Adaptive complexity and phenomenal consciousness. *Philosophy Of Science* 67:648-670.

Ornstein, R. 1991. *The Evolution of Consciousness: Of Darwin, Freud, and Cranial Fire: The Origins of the Way We Think*. Prentice-Hall.

Povinelli, D.J. 1987. Monkeys, apes, mirrors, minds: The evolution of self-awareness in primates. *Human Evolution* 2:493-507.

Pribram, K.H., Jerison, H.J., McGuinness, D., & Eccles, J.C. 1982. The evolution of consciousness: A symposium. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Reber, A.S. 1992. The cognitive unconscious: An evolutionary perspective. *Consciousness and Cognition* 1:93-133.

Reber, A.S. 1992. An evolutionary context for the cognitive unconscious. *Philosophical Psychology* 5:33-51.

Reber, A. & Allen, R. 2000. Individual differences in implicit learning: Implications for the evolution of consciousness. In (R. Kunzendorf & B. Wallace, eds) *Individual differences in conscious experience*. John Benjamins.

Rogers, L.J. 1995. Evolution and development of brain asymmetry, and its relevance to language, tool use and consciousness. *International Journal of Comparative Psychology* 8:1-15.

Roth, G. 2000. The evolution and ontogeny of consciousness. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Towers, B. 1979. Consciousness and the brain: Evolutionary aspects. In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.

Vandervert, L.R. 1995. Chaos theory and the evolution of consciousness and mind: A thermodynamic/holographic resolution to the mind-body problem. *New Ideas in Psychology* 13:107-27.

6.4b Consciousness and Language

Arbib, M.A. 1972. Consciousness: The secondary role of language. *Journal of Philosophy* 69.

Bailey, W. 1986. Consciousness and action/motion theories of communication. *Western Journal of Speech Communication* 50:74-86.

Blachowicz, J. 1997. The dialogue of the soul with itself. *Journal of Consciousness Studies* 4:485-508.

Carruthers, P. 1996. The involvement of language in conscious thinking. In *Language, Thought, and Consciousness*. Cambridge University Press.

Chafe, W.L. 1980. The deployment of consciousness in the construction of narrative. In (W. Chafe, ed) *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex.

Chafe, W.L. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.

Chafe, W.L. 1996. How consciousness shapes language. *Pragmatics and Cognition* 4:35-54.

Chafe, W. 2000. A linguist's perspective on William James and "The Stream of Thought." *Consciousness & Cognition* 9:618-628.

Chapman, S.B. & Ulatowska, H.K. 1997. Discourse in dementia: Considerations of Consciousness. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

de Beaugrande, R. 1997. The "conscious and unconscious mind" in the theoretical discourse of modern linguistics. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

Fludernik, M., & Sell, R.D. 1995. The fictions of language and the languages of fiction: The linguistic representation of speech and consciousness. *Journal of Pragmatics* 24:557.

Johnston, P.K. 1997. Battle within: Shakespeare's brain and the nature of human consciousness. *Journal of Consciousness Studies* 4:365-73.

Langacker, R.W. 1997. Consciousness, construal, and subjectivity. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

Lecours, A.R. 1998. Language contrivance on consciousness (and vice versa). In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Macphail, E. 2000. The search for a mental Rubicon. In (C. Heyes & L. Huber, eds) *_The Evolution of Cognition_*. MIT Press.

Markey, J.F. 1925. The place of language habits in a behavioristic explanation of consciousness. *Psychological Review* 32:384-401.

Pronko, N.H. 1987. Language with or without consciousness. In (G. Greenberg & E. Tobach, eds) *_Cognition, Language and Consciousness: Integrative Levels_*. Lawrence Erlbaum.

Ricciardelli, L.A. 1993. Two components of metalinguistic awareness: Control of linguistic processing and analysis of linguistic knowledge. *Applied Psycholinguistics* 14:349-367.

Schooler, J.W & Fiore, S.M. 1997. Consciousness and the limits of language: You can't always say what you think or think what you say. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Sekhar, A.C. 1948. Language and consciousness. *Indian Journal of Psychology* 23:79-84.

Sinha, V. 1987. Symbolic language not a pre-requisite for self-awareness. *Psycho-Lingua* 17:115-121.

Stamenov, M.I. (ed) 1997. *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

Stamenov, M.I. 1997. Grammar, meaning, and consciousness: What sentence structure can tell us about the structure of consciousness. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

6.4c Animal Consciousness

- Allen, G.E. 1987. Materialism and reductionism in the study of animal consciousness. In (G. Greenberg, E. Tobach, eds) *Cognition, Language, and Consciousness: Integrative Levels*. Lawrence Erlbaum.
- Bekoff, M. 1992. Scientific ideology, animal consciousness, and animal protection: A principled plea for unabashed common sense. *New Ideas in Psychology* 10:79-94.
- Bradshaw, R.H. 1998. Consciousness in nonhuman animals: Adopting the precautionary principle. *Journal of Consciousness Studies* 5:108-14.
- Burghardt, G. 1985. Animal awareness: Current perceptions and historical perspective. *American Psychologist* 40:905-919.
- Carruthers, P. 1989. Brute experience. *Journal of Philosophy* 258-69.
- Cheney, D.L. & Seyfarth, R.M. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. University of Chicago Press.
- Crook, J.H. 1983. On attributing consciousness to animals. *Nature* 303:11-14.
- Dawkins, M.S. 1993. *Through Our Eyes Only: The Search for Animal Consciousness*.
- Denett, D.C. 1995. Animal consciousness: What matters and why? *Social Research* 62:691-710.
- Eccles J.C. 1982. Animal consciousness and human self-consciousness. *Experientia* 38:1384-91.
- Gallup, G.G. 1985. Do minds exist in species other than our own? *Neuroscience and Biobehavioral Reviews* 9:631-41.
- Griffin, D.R. 1981. *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience*. William Kaufmann.
- Griffin, D.R. 1985. Animal consciousness. *Neuroscience and Biobehavioral Reviews* 9:615-22.
- Griffin, D.R. 1992. *Animal Minds*. University of Chicago Press.
- Griffin, D.R. 1995. Windows on animal minds. *Consciousness and Cognition* 4:194-204.
- Heyes, C.M. 1987. Cognisance of consciousness in the study of animal knowledge. In (W. Callebaut & R. Pinxten, eds) *Evolutionary Epistemology: A Multiparadigm Program*. Reidel.
- Hughes, H. 2001. *Sensory Exotica: A World Beyond Human Experience*. MIT Press.
- Jolley, N. 1995. Sensation, intentionality, and animal consciousness. *Ratio* 8:128-42.
- Jolly, A. 1991. Conscious chimpanzees? A review of recent literature. In (C. Ristau, ed) *Cognitive Ethology*. Lawrence Erlbaum.
- Latto, R. 1986. The question of animal consciousness. *Psychological Record* 36:309-14.
- Oakley, D.A. 1985. Animal awareness, consciousness, and self-image. In (D. Oakley, ed) *Brain and Mind*. Methuen.

Radner, D. & Radner, M. 1996. Animal Consciousness. Prometheus Books.

Reiss, D. 1998. Cognition and communication in dolphins: A question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness II. MIT Press.

Ristau, C.A. 1983. Language, cognition, and awareness in animals? *Annals of the New York Academy of Sciences* 406:170-86.

Roberts, H. 1968. Consciousness in animals and automata. *Psychological Reports* 22:1226-28.

Rollin, B.E. 1986. Animal consciousness and scientific change. *New Ideas in Psychology* 4:141-52.

Rollin, B.E. 1989. The Unheeded Cry: Animal Consciousness, Animal Pain, and Science. Oxford University Press.

Rothschild, M. 1993. Thinking about animal consciousness. *Journal of Natural History* 27:509-12.

Rushen, J.P. 1985. The scientific status of animal consciousness. *Applied Animal Behaviour Science* 13:387-390.

Savage-Rumbaugh, E.S. & Rumbaugh, D. 1998. Perspectives on consciousness, language, and other emergent processes in apes and humans. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness II. MIT Press.

Tye, M. 1997. The problem of simple minds: Is there anything it's like to be a honeybee? *Philosophical Studies* 88:289-317.

van Rooijen, J. 1981. Are feelings adaptations? The basis of modern applied animal ethology. *Applied Animal Ethology* 7:187-89.

Weiskrantz, L. 1995. The problem of animal consciousness in relation to neuropsychology. *Behavioral Brain Research* 71:171-75.

6.4d Animal Self-Consciousness

Byrne, R.W. & Whiten, A. 1988. Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans. Oxford University Press.

Epstein, R., Lanza, R.P. & Skinner, B.F. 1981. "Self-awareness" in the pigeon. *Science* 212:695-96.

Gallup, G.G. 1970. Chimpanzees: Self-recognition. *Science* 167:86-87.

Gallup, G.G. 1975. Toward an operational definition of self-awareness. In (R. Tuttle, ed) Socioecology and the Psychology of Primates. Mouton.

Gallup, G.G. 1977. Self-recognition in primates: A comparative approach to the bidirectional properties of consciousness. *American Psychologist* 32:329-38.

Gallup, G.G. 1979. Self-recognition in chimpanzees and man: A developmental and comparative perspective. In (M. Lewis & M. Rosenblum, eds) Genesis of Behavior, Volume 2. Plenum Press.

Gallup, G.G. 1982. Self-awareness and the emergence of mind in primates. *American Journal of Primatology* 2:237-48.

Gallup, G.G. 1987. Self-awareness. In (G. Mitchell, ed) Comparative Primate

Biology, Volume 2_. Liss.

Gallup, G.G. 1991. Toward a comparative psychology of self-awareness: Species limitations and cognitive consequences. In (G. Goethals & J. Strauss, eds) The Self: An Interdisciplinary Perspective. Springer-Verlag.

Gallup, G.G. 1994. Self-recognition: Research strategies and experimental design. In (S. Parker, R. Mitchell, & M. Boccia, eds) Self-Awareness in Animals and Humans: Developmental Perspectives. Cambridge University Press.

Hart, D. & Karmel, M.P. 1996. Self-awareness and self-knowledge in humans, apes, and monkeys. In (A. Russon, K. Bard, & S. Parkers, eds) Reaching into Thought: The Minds of the Great Apes. Cambridge University Press.

Heyes, C.M. 1994. Reflections on self-recognition in primates. *Animal Behaviour* 47:909-19.

Hyatt, C.W. & Hopkins, W. 1994. Self-awareness in bonobos and chimpanzees: A comparative perspective. In (S. Parker, R. Mitchell, & M. Boccia, eds) Self-Awareness in Animals and Humans: Developmental Perspectives. Cambridge University Press.

Marten, K. & Psarakos, S. 1992. Using self-view television to distinguish between self-examination and social behavior in the bottlenose dolphin. *Consciousness and Cognition* 4:205-24.

Marten, K. & Psarakos, S. 1994. Evidence for self-awareness in the bottlenose dolphin. In (S. Parker, R. Mitchell, & M. Boccia, eds) Self-Awareness in Animals and Humans: Developmental Perspectives. Cambridge University Press.

Miles, H.L. 1994. Me Chantek: The development of self-awareness in a signing orangutan. In (S. Parker, R. Mitchell, & M. Boccia, eds) Self-Awareness in Animals and Humans: Developmental Perspectives. Cambridge University Press.

Moynihan, M.H. 1997. Self-awareness, with specific references to Coleoid cephalopods. In (R. Mitchell, N. Thompson, & H. Miles, eds) Anthropomorphism, Anecdotes, and Animals. SUNY Press.

Parker, S.T. 1991. A developmental approach to the origins of self-recognition in great apes. *Human Evolution* 6:435-49.

Patterson, F.G.P. & Cohn, R. 1994. Self-recognition and self-awareness in lowland gorillas. In (S. Parker, R. Mitchell, & M. Boccia, eds) Self-Awareness in Animals and Humans: Developmental Perspectives. Cambridge University Press.

Povinelli, D.J. 1987. Monkeys, apes, mirrors, minds: The evolution of self-awareness in primates. *Human Evolution* 2:493-507.

Suarez, S.D. & Gallup, G.G. 1981. Self-recognition in chimpanzee and orangutans, but not gorillas. *Journal of Human Evolution* 10:175-88.

Swartz, K.B. & Evans, S. 1991. Not all chimpanzees show self-recognition. *Primates* 32:483-96.

6.4e Altered States of Consciousness

Atkinson, R.P. & H. Earl. 1996. Enhanced vigilance in guided meditation: Implications of altered consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness. MIT Press.

Austin, J.H. 1998. Zen and the Brain: Toward an Understanding of Meditation

and Consciousness_. MIT Press.

Forman, R. (ed) 1990. _The Problem of Pure Consciousness: Mysticism and Philosophy_. Oxford University Press.

Forman, R. 1998. What does mysticism have to teach us about consciousness? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness 1996_. MIT Press.

Goleman, D. 1976. Meditation and consciousness: An Asian approach to mental health. American Journal of Psychotherapy 30:41-54.

Hilgard, E.R. 1979. Consciousness and control: Lessons from hypnosis. Australian Journal of Clinical & Experimental Hypnosis 7:103-15.

Hunt H.T. 1985. Cognition and states of consciousness: the necessity for empirical study of ordinary and nonordinary consciousness for contemporary cognitive psychology. Perceptual and Motor Skills 60:239-82.

Katz, J.M. 1983. Altered states of consciousness and emotion. Imagination, Cognition and Personality 2:37-50.

Keen, E. 2000. _Chemicals for the Mind: Psychopharmacology and Human Consciousness_. Praeger.

Kunzendorf, R.G., Beltz, S.M. & Tymowicz, G. 1992. Self-awareness in autistic subjects and deeply hypnotized subjects: Dissociation of self-concept versus self-consciousness. Imagination, Cognition and Personality 11:129-41.

Munglani, R. & Jones, J.G. 1992. Sleep and general anesthesia as altered states of consciousness. Journal of Psychopharmacology 6:399-409.

Novak, P. 1996. Buddhist meditation and consciousness of time. Journal of Consciousness Studies 3:267-77.

Oakley, D. 1999. Hypnosis and consciousness: A structural model. Contemporary Hypnosis 16:215-223.

Pekala, R.J. & Kumar, V.K. 1989. Phenomenological patterns of consciousness during hypnosis: Relevance to cognition and individual differences. Australian Journal of Clinical and Experimental Hypnosis 17:1-20.

Pekala, R.J. & Cardena, E. 2000. Methodological issues in the study of altered states of consciousness and anomalous experiences. In (E. Cardena & S. Lynn, eds) _Varieties of Anomalous Experience: Examining the Scientific Evidence_. American Psychological Association.

Pekala, R. & Kumar, V. 2000. Individual differences in patterns of hypnotic experience across low and high hypnotically susceptible individuals. In (R. Kunzendorf & B. Wallace, eds) _Individual Differences in Conscious Experience. John Benjamins.

Shapiro, D.H. 1982. Meditation as an altered state of consciousness: Contributions of Western behavioral science. Journal of Transpersonal Psychology 15:61-81.

Spivak, L., V. Puzenko, S. Medvedev, & Y. Polyakov 1990. Neurophysiological correlates of the altered state of consciousness during hypnosis. Human Physiology 16:405-410.

Tart, C.T. (ed) 1990. _Altered States of Consciousness_ (third edition). Harper Collins.

Tart, C.T. 1998. Transpersonal psychology and methodologies for a

comprehensive science of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Tart, C.T. 2000. Investigating altered states of consciousness on their own terms: State-specific sciences. In (M. Velmans, eds) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

Tinnin, L. 1990. Mental unity, altered states of consciousness, and dissociation. *Dissociation: Progress in the Dissociative Disorders* 3:154-59.

Travis, F. & Pearson, C. 2000. Pure consciousness: Distinct phenomenological and physiological correlates of "consciousness itself". *International Journal of Neuroscience* 100:77-89.

West, M. 1983. Meditation and self-awareness: Physiological and phenomenological approaches. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Wolman, B.B. & Ullman, U. 1986. *Handbook of States of Consciousness*. van Nostrand Reinhold.

Venkatesh S., Raju T.R., Shivani, Y., Tompkins G., & Meti B.L. 1997. A study of structure of phenomenology of consciousness in meditative and non-meditative states. *Indian Journal of Physiology and Pharmacology* 41:149-53.

Walsh, R. 1998. States and stages of consciousness: Current research and understanding. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

6.4f Parapsychology and Consciousness

Beloff, J. 1973. The subliminal and the extrasensory. *Parapsychology Review* 4:23-27.

Beloff, J. 1976. Mind-body interactionism in light of the parapsychological evidence. *Theoria to Theory* 10:125-37.

Beloff, J. 1980. Could there be a physical explanation for psi? *Journal of the Society for Psychical Research* 50:263-272

Beloff, J. 1987. Parapsychology and the mind-body problem. *Inquiry* 30:215-25.

Beloff, J. 1989. Dualism: A parapsychological perspective. In (J. Smythies & J. Beloff, eds) *The Case for Dualism*. Virginia University Press.

Bem, D.J. & Honorton, C. 1994. Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin* 115:4-18.

Bierman, D. 1998. Do psi phenomena suggest radical dualism? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Blackmore, S. 1991. Psi in science. *Journal of the Society for Psychical Research* 57:404-11.

Blackmore, S. 1998. Why psi tells us nothing about consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Bohm, D.J. 1986. A new theory of the relationship of mind and matter. *Journal of the American Society for Psychical Research* 80:113-35.

Braud, W.G. 1994. The role of mind in the physical world: A psychologist's

view. *European Journal of Parapsychology* 10:66-77.

Braude, S.E. 1979. *_ESP and Psychokineses: A Philosophical Examination_*. Temple University Press.

Braude, S.E. 1986. *_The Limits of Influence: Psychokinesis and the Philosophy of Science_*. Routledge and Kegan Paul.

Burns, J.E. 1993. Current hypotheses about the nature of the mind-brain relationship and their relationship to findings in parapsychology. In (K. Rao, ed) *_Cultivating Consciousness_*. Praeger.

Burns, J.E. 1993. Time, consciousness, and psi. In (B. Kane, J. Millay, & D.H. Brown, eds) *_Silver Threads: 25 Years of Parapsychology Research_*. Praeger.

Burns, J.E. 1986. Consciousness and psi. *PSI Research* 5:166-205.

Dilley, F.B. 1989. Mind-brain interaction and psi. *Southern Journal of Philosophy* 26:469-80.

Dilley, F.B. 1990. Telepathy and mind-brain dualism. *Journal of the Society for Psychical Research* 56:129-37.

Edge, H.L. 1989. Psi, self, and the new mentalism. In (L. Henkel & J. Palmer, eds) *_Research in Parapsychology 1989_*. Scarecrow Press.

Eisenbud, J. 1975. The mind-matter interface. *Journal of the American Society for Psychical Research* 69:115-26.

Goswami, A. 1986. The quantum theory of consciousness and psi. *PSI Research* 5:145-65.

Griffin, D.R. 1993. Parapsychology and philosophy: A Whiteheadian postmodern perspective. *Journal of the American Society for Psychical Research* 87:217-88.

Griffin, D.R. 1994. Dualism, materialism, idealism, and psi: A reply to John Palmer. *Journal of the American Society of Psychical Research* 88:23-39.

Grof, S. 2000. *_Psychology of the Future: Lessons from Modern Consciousness Research_*. State University of New York Press.

Heath, P. 2000. The PK zone: A phenomenological study. *Journal of Parapsychology* 64:53-72.

Honorton, C. 1985. Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology* 1:51-91.

Hubbard, T.L. 1996. Consciousness and cognition beyond the body: Functionalist cognitive science and the possibility of out-of-body experiences and reincarnation. *Journal of the American Society for Psychical Research* 90:202-20.

Hyman, R. 1985. The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology* 49:3-49.

Hyman, Ray & Honorton, C. 1986. A joint communique: The psi ganzfeld controversy. *Journal of Parapsychology* 50:351-64.

Hyman, R. 1994. Anomaly or artifact? Comments on Bem and Honorto. *Psychological Bulletin* 115:19-24.

Jahn, R.G. & Dunne, B.J. 1987. *_Margins of Reality: The Role of Consciousness in the Physical World_*. Harcourt Brace Jovanovich.

- Kreitler, H. & Kreitler, S. 1973. Subliminal perception and extrasensory perception. *Journal of Parapsychology* 37:163-88.
- Krippner, S. & George, L. 1986. Psi phenomena as related to altered states of consciousness. In (B. Wolman & M. Ullman, eds) *Handbook of States of Consciousness*. van Nostrand Reinhold.
- Mattuck, R. 1982. A crude model of the mind-matter interaction using Bohm-Bub hidden variables. *Journal of the Society for Psychical Research* 51:238-245.
- Nash, C.B. 1976. Psi and the mind-body problem. *Journal of the Society for Psychical Research* 48:267-70.
- Nash, C.B. 1995. A panpsychic theory of mind and matter. *Journal of the Society for Psychical Research* 60:171-73.
- Poynton, J.C. 1994. Making sense of psi: Whitehead's multilevel ontology. *Journal of the Society for Psychical Research* 59:401-12.
- Price, E.A. 1981. A "three worlds" perspective to the mind-brain relationship in parapsychology. *Parapsychological Journal of South Africa* 2:38-49.
- Rao, K.R. & Palmer, J. 1987. The anomaly called psi: Recent research and criticism. *Behavioral and Brain Sciences* 10:539-51.
- Rao, K.R. 1991. Consciousness research and psi. *Journal of Parapsychology* 55:1-43.
- Rauscher, E.A. 1983. Multidimensional properties of consciousness and some laws of reality. *PSI Research* 2:53-66.
- Richards, D.G. 1996. Psi and the spectrum of consciousness. *Journal of the American Society for Psychical Research* 90:251-67.
- Roberts, F.S. 1991. Some apparently non-cerebral aspects of consciousness. *Journal of the Society for Psychical Research* 58:31-38.
- Roberts, F.S. 1995. Is physically-based consciousness a reality? *Journal of the Society for Psychical Research* 60:398-400.
- Roney-Dougal, S.M. 1986. Subliminal and psi perception: A review of the literature. *Journal of the Society for Psychical Research* 53:405-34.
- Smythies, J.M. 1960. Three classical theories of mind. *Journal of the Society for Psychical Research* 40:385-397.
- Stokes, D.M. 1982. On the relationship between mind and brain. *Parapsychology Review*, 13:22-27.
- Stokes, D.M. 1993. Mind, matter, and death: Cognitive neuroscience and the problem of survival. *Journal of the American Society for Psychical Research* 87:41-84.
- Stokes, D.M. 1997. *The Nature of Mind: Parapsychology and the Role of Consciousness in the Physical World*. McFarland and Co.
- Tiller, W., Kohane, M., & Dibble, W. 2000. Can an aspect of consciousness be imprinted into an electronic device? *Integrative Physiological & Behavioral Science* 35:142-163.
- Varvoglis, M. 1996. Nonlocality on a human scale: Psi and consciousness research. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Wade, J. 1998. Physically transcendent awareness: A comparison of the phenomenology of consciousness before birth and after death. *Journal of Near-Death Studies* 16:249-275.

Walker, E.H. 1984. A review of criticisms of the quantum-mechanical theory of psi phenomena. *Journal of Parapsychology* 48:277-32.

Woodworth, H. 1942. Report of investigations into an obscure function of the subconscious mind. *Journal of the American Society for Psychical Research* 36:185-230.

6.4g Phenomenology

Ackerman, D. 1990. *_A Natural History of the Senses_*. Randhom House,

Arvidson, P.S. 1992. On the origin of organization in consciousness. *Journal of the British Society of Phenomenology* 23:53-65.

Arvidson, P.S. 1996. Toward a phenomenology of attention. *Human Studies* 19:71-84.

Baars, B.J. 1993. Putting the focus on the fringe: Three empirical cases. *Journal of Consciousness Studies* 2:126-36.

Chokr, N.N. 1992. Mind, consciousness, and cognition: Phenomenology vs. cognitive science. *Husserl Studies* 9:179-97.

Deikman, A. 1996. 'I' = awareness. *Journal of Consciousness Studies* 3:350-56.

Depraz, N., Varela, F., & Vermersch, P. 2000. The gesture of awareness: An account of its structural dynamics. In (M. Velmans, ed) *_Investigating Phenomenal Consciousness: New Methodologies and Maps_*. John Benjamins.

de Quincey, C. 2000. Intersubjectivity: Exploring consciousness from the second-person perspective. *Journal of Transpersonal Psychology* 32:135-155.

Diaz, J. 1996. The stream revisited: A process model of phenomenological consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Ellis, R. 1986. *_An Ontology of Consciousness_*. Kluwer.

Galin, D. 1994. The structure of awareness: Contemporary applications of William James' forgotten concept of "the fringe". *Journal of Mind and Behavior* 15:375-401.

Galin, D. 1996. The structure of subjective experience: Sharpen the concepts and terminology. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Gallagher, S. 1997. Mutual enlightenment: Recent phenomenology in cognitive science. *Journal of Consciousness Studies* 4:195-214.

Gurwitsch, A. 1964. *_The Field of Consciousness_*. Duquesne University Press.

Gurwitsch, A. 1966. *_Studies in Phenomenology and Psychology_*. Northwestern University Press.

Ihde, D. 1977. *_Experimental Phenomenology_*. Putnam.

Jopling, D.A. 1996. Sub-phenomenology. *Human Studies* 19:153-73.

Koestenbaum, P. 1962. The sense of subjectivity. Review of Existential Psychology 2:47-65.

Lind, R. 1996. Micro-phenomenology: Toward a hypothetico-inductive science of experience. International Philosophical Quarterly 36:429-42.

Mangan, B. 1993. Taking phenomenology seriously: The "fringe" and its implication for cognitive research. Consciousness and Cognition 2:89-108.

Marbach, E. 1993. _Mental Representation and Consciousness: Toward a Phenomenological Theory of Representation and Reference_. Kluwer.

Marbach, E. 1996. Understanding the representational mind: A phenomenological perspective. Human Studies 19:137-52.

Marbach, E. 2000. The place for an Ego in current research. In (D. Zahavi, ed) _Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience_. John Benjamins.

Natsoulas, T. 1997. The presence of environmental objects to perceptual consciousness: An integrative, ecological and phenomenological approach. Journal of Mind & Behavior 18:371-390.

Natsoulas, T. 1997. The presence of environmental objects to perceptual consciousness: A difference it makes for psychological functioning. American Journal of Psychology 110:507-526.

Nelson, P. 1998. Consciousness as reflexive shadow: An operational psychophenomenological model. Imagination, Cognition and Personality 17:215-228.

Pekala, R.J., Wenger C.F., & Levine R.L. 1985. Individual differences in phenomenological experience: states of consciousness as a function of absorption. Journal of Personality and Social Psychology 48:125-32.

Pekala, R.J. & Levine, R.L. 1982. Mapping consciousness: Development of an empirical-phenomenological approach. Imagination, Cognition & Personality 1:29-47.

Rao, K.R. 1998. Two faces of consciousness: A look at Eastern and Western perspectives. Journal of Consciousness Studies 5:309-27.

Shanon, B. 1984. The case for introspection. Cognition and Brain Theory 7:167-80.

Shear, J. 1996. The hard problem: Closing the empirical gap. Journal of Consciousness Studies 3:54-68. Reprinted in (J. Shear, ed) _Explaining Consciousness: The Hard Problem_. MIT Press,

Stevens, R. 2000. Phenomenological approaches to the study of conscious awareness. In (M. Velmans, ed) _Investigating Phenomenal Consciousness: New Methodologies and Maps_. John Benjamins.

Wilber, K. 2000. Waves, streams, states and self: Further considerations for an integral theory of consciousness. Journal of Consciousness Studies 7:145-176.

Varela, F. 1995. Neurophenomenology: A methodological remedy for the hard problem. Journal of Consciousness Studies 3:330-49. Reprinted in (J. Shear, ed) _Explaining Consciousness: The Hard Problem_. MIT Press.

6.4h Foundations

Baars, B.J. 1994. A thoroughly empirical approach to consciousness. Psyche 1.

Baruss, I., & Moore, R.J. 1992. Measurement of beliefs about consciousness and reality. *Psychological Reports* 71:59-64.

Battista, J.R. 1978. The science of consciousness. In (K.S. Pope & J.L. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Block, N. 2001. Paradox and cross purposes in recent work on consciousness. *Cognition* 79:197-219.

Conrad, D. 1996. Consciousness, privacy, and information. *Biosystems* 38:207-10.

Dennett, D. 2001. Are we explaining consciousness yet? *Cognition* 79:221-37.

Dunlop, K. 1912. The case against introspection. *Psychological Review* 19:404-13.

Flanagan, O.J. 1995. Consciousness and the natural method. *Neuropsychologia* 33:1103-15.

Foss, J. 2000. *Science and the Riddle of Consciousness: A Solution*. Kluwer Academic Publishers.

Goldman, A. 1997. Science, publicity, and consciousness. *Philosophy of Science* 64:525-45.

Goldman, A. 2000. Can science know when you're conscious? Epistemological foundations of consciousness research. *Journal Of Consciousness Studies* 7:3-22.

Grinker, R.R. 1953. Problems of consciousness: A review, an analysis, and a proposition. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Fourth Conference*. Josiah Macy Foundation.

Jack, A.I. & Shallice, T. 2001. Introspective physicalism as an approach to the science of consciousness. *Cognition* 79:161-196.

Lyons, W. 1986. *The Disappearance of Introspection*. MIT Press.

Miller, D. 2000. Designing a bridge for consciousness: Are criteria for a unification of approaches feasible? *Advances in Mind-Body Medicine* 16:82-89.

Nunez, R. 1997. Eating soup with chopsticks: Dogmas, difficulties, and alternatives in the study of conscious experience. *Journal of Consciousness Studies* 4:143-66.

Pekala, R. & Cardena, E. 2000. Methodological issues in the study of altered states of consciousness and anomalous experiences. In (E. Cardena & S. Lynn, eds) *Varieties of Anomalous Experience: Examining the Scientific Evidence*. American Psychological Association.

Scott, A.C. 1998. Reductionism revisited. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Scott, A. 2000. Modern science and the mind. In (M. Velmans, ed) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

Searle, J. 1998. How to study consciousness scientifically. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Simon, H.A. 1997. Scientific approaches to the question of consciousness. In

(J. Cohen & J. Schooler, eds) _Scientific Approaches to Consciousness_.
Lawrence Erlbaum.

Stevens, S.S. 1966. Quantifying the sensory experience. In (P. Feyerabend & G. Maxwell, eds) _Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl_. University of Minnesota Press.

Varela, F. 1998. A science of consciousness as if experience mattered. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness 1996_. MIT Press.

Velmans, M. 1994. A reflexive science of consciousness. In _Experimental and Theoretical Studies of Consciousness_ (Ciba Foundation Symposium 174). Wiley.

Velmans, M. 1996. Introduction to the science of consciousness. In (M. Velmans, ed) _The Science of Consciousness_. Routledge.

Velmans, M. 1998. Goodbye to reductionism: Complementary first and third-person approaches to consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness II_. MIT Press.

Velmans, M. 2000. _Understanding Consciousness_. Routledge.

Wallace, B. 2000. _The Taboo of Subjectivity: Toward a New Science of Consciousness_. Oxford University Press.

Williams, D.C. 1934. Scientific method and the existence of consciousness. *Psychological Review* 41:461-79.

6.4i Consciousness and Science, Misc

Abramson, H.A. (ed) 1950. _Problems of Consciousness: Transactions of the First Conference_. Josiah Macy Foundation.

Abramson, H.A. (ed) 1951. _Problems of Consciousness: Transactions of the Second Conference_. Josiah Macy Foundation.

Abramson, H.A. (ed) 1952. _Problems of Consciousness: Transactions of the Third Conference_. Josiah Macy Foundation.

Abramson, H.A. (ed) 1953. _Problems of Consciousness: Transactions of the Fourth Conference_. Josiah Macy Foundation.

Abramson, H.A. (ed) 1954. _Problems of Consciousness: Transactions of the Fifth Conference_. Josiah Macy Foundation.

Bielecki, A., Kokoszka, A., & Holas, P. 2000. Dynamic systems theory approach to consciousness. *International Journal of Neuroscience* 104:29-47.

Blakemore, C. & Greenfield, S. 1987. _Mindwaves: Thoughts on Intelligence, Identity, and Consciousness_. Blackwell.

Bock, G.R. & Marsh, J. (eds) 1993. _Experimental and Theoretical Studies of Consciousness_ (Ciba Foundation Symposium 174). Wiley.

Cohen, J.D. & Schooler, J.W. (eds) 1997. _Scientific Approaches to Consciousness_. Lawrence Erlbaum.

Cornwell, J. (ed) 1998. _Consciousness and Human Identity_. Oxford University Press.

Cotterill, R. 2000. _Enchanted Looms: Conscious Networks in Brains and Computers_. Cambridge University Press.

Hameroff, S.R, Kaszniak, A. & Scott, A. (eds) 1996. *_Toward a Science of Consciousness: The First Tucson Discussions and Debates_*. MIT Press.

Ito, M., Miyashita, Y., & Rolls, E.T. (eds) 1997. *_Cognition, Computation, and Consciousness_*. Oxford University Press.

Jarvilehto, T. 2000. The theory of the organism-environment system: The problem on mental activity and consciousness. *Integrative Physiological & Behavioral Science* 35:35-57.

John, E.R. 2001. A field theory of consciousness. *Consciousness and Cognition* 10:184-213.

Josephson, B. & Ramachandran, V.S. (eds) 1980. *_Consciousness and the Physical World_*. Pergamon Press.

Keyes, C. D. 1999. *_Brain Mystery Light and Dark: The Rhythm and Harmony of Consciousness_*. Routledge.

Marcel, A.J. & Bisiach, E. (eds) 1988. *_Consciousness in Contemporary Science_*. Oxford University Press.

Oakley, D.A. (ed) 1985. *_Brain and Mind_*. Methuen.

Scott, A. 1995. *_Stairway to the Mind: The Controversial New Science of Consciousness_*. Springer.

Sugarman, A.A. & Tarter, R.E. (eds) 1978. *_Expanding Dimensions of Consciousness_*. Springer.

Torey, Z. 1999. *_The Crucible of Consciousness_*. Oxford University Press.

Velmans, M. (eds) 1996. *_The Science of Consciousness: Tutorial Essays_*. Routledge.

Wilber, K. 2000. *_Integral Psychology: Consciousness, Spirit, Psychology, Therapy_*. Shambhala.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 1: Consciousness and Qualia [1082]

=====

- 1.1 Consciousness [85]
 - 1.1a General [47]
 - 1.1b The Concept of Consciousness [38]
- 1.2 Explaining Consciousness? [115]
 - 1.2a Subjectivity and Objectivity (Nagel) [46]
 - 1.2b The Explanatory Gap (Levine) [16]
 - 1.2c 'Hard' and 'Easy' Problems (Chalmers) [25]
 - 1.2d Cognitive Closure (McGinn) [16]
 - 1.2e Miscellaneous [12]
- 1.3 Materialism and Dualism [213]
 - 1.3a The Knowledge Argument (Jackson) [46]
 - 1.3b Zombies & Modal Arguments [38]
 - 1.3c Essentialist Arguments (Kripke) [31]
 - 1.3d Arguments from Disembodiment [17]
 - 1.3e Consciousness and Physicalism, Misc [14]
 - 1.3f Consciousness and Dualism [16]
 - 1.3g Mind-Body Problem, Misc [51]
- 1.4 Specific Views on Consciousness [227]
 - 1.4a Higher-Order Thought Approaches (Rosenthal, etc) [46]
 - 1.4b Dennett on Consciousness [60]
 - 1.4c Functionalism [8]
 - 1.4d Eliminativism [16]
 - 1.4e Epiphenomenalism [30]
 - 1.4f Interactionism [16]
 - 1.4g Panpsychism [24]
 - 1.4h Intrinsic Monism (Russell, Feigl, etc) [27]
- 1.5 Consciousness and Content [129]
 - 1.5a Consciousness and Intentionality (Searle, etc) [23]
 - 1.5b The Content of Experience [48]
 - 1.5c Representationalism [38]
 - 1.5d Internalism and Externalism about Experience [12]
 - 1.5e Miscellaneous [8]
- 1.6 Aspects of Consciousness [182]
 - 1.6a Self-Consciousness [54]
 - 1.6b Unity of Consciousness [45]
 - 1.6c Homogeneity of Consciousness (Sellars, etc) [11]
 - 1.6d Knowledge of Consciousness [50]
 - 1.6e The Function of Consciousness [22]
- 1.7 Qualia [131]
 - 1.7a Qualia, General [27]
 - 1.7b Qualia and Materialism [13]
 - 1.7c Eliminativism about Qualia [11]
 - 1.7d The Inverted Spectrum [31]
 - 1.7e Absent Qualia (Block, etc) [15]
 - 1.7f Introspection and Absent Qualia (Shoemaker) [10]
 - 1.7g Functionalism and Qualia, General [24]

1.1 Consciousness

1.1a Consciousness -- General

Armstrong, D.M. & Malcolm, N. 1984. Consciousness and Causality: A Debate on the Nature of Mind. Blackwell.

Block, N., Flanagan, O., & Guzeldere, G. (eds) 1997. The Nature of Consciousness: Philosophical Debates. MIT Press.

An anthology of central philosophical papers on consciousness.

Carruthers, P. 2000. _Phenomenal Consciousness: A Naturalistic Theory_.
Cambridge University Press.

Catalano, J. 2000. _Thinking Matter: Consciousness from Aristotle to Putnam and Sartre_. Routledge.

Chalmers, D.J. 1991. Consciousness and cognition. Manuscript.
Exploring the link between consciousness and judgments about consciousness.
Coherence between these => consciousness depends on the functional but isn't
reducible. Toward a dual-aspect theory based on pattern and information.

Chalmers, D.J. 1996. _The Conscious Mind: In Search of a Fundamental Theory_.
Oxford University Press.
Argues against the reductive explanation of consciousness, and for a kind of
naturalistic dualism. Moves toward a "fundamental theory" to bridge the
gap, and draws out some consequences.

Churchland, P.M. & Churchland, P.S. 1997. Recent work on consciousness:
Philosophical, theoretical, and empirical. *Seminars in Neurology* 17:179-86.

Davies, M. & Humphreys, G. 1993. _Consciousness: Philosophical and
Psychological Essays_. Blackwell.
A collection of 5 psychological and 8 philosophical essays on consciousness.

Flanagan, O.J. 1991. Consciousness. In _The Science of the Mind_. MIT Press.
On the mysteries of consciousness. Argues with epiphenomenalism, "conscious
inessentialism", and the "new mysterians" (Nagel, McGinn). Toward a
naturalistic theory, drawing on ideas of Edelman, Calvin, Dennett.

Flanagan, O.J. 1992. _Consciousness Reconsidered_. MIT Press.
Argues that consciousness can be accounted for in a naturalistic framework.
With arguments against eliminativism and epiphenomenalism, evidence from
neuroscience and psychology, and discussions of the stream and the self.

Flanagan, O.J. & Guzeldere, G. 1997. Consciousness: A philosophical tour.
In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) _Cognition, Computation, and
Consciousness_. Oxford University Press.

Foss, J. 2000. _Science and the Riddle of Consciousness: A Solution_. Kluwer.

Graham, G. & Horgan, T. 1998. Sensations and grain processes. In (G.
Mulhauser, ed) _Evolving Consciousness_. John Benjamins.

Gregory, R.L. 1988. Consciousness in science and philosophy: conscience and
con-science. In (A. Marcel & E. Bisiach, eds) _Consciousness in Contemporary
Science_. Oxford University Press.

Guzeldere, G. 1995. Consciousness: what it is, how to study it, what to learn
from its history. *Journal of Consciousness Studies* 2:30-51.
A history of the study of consciousness, especially in psychology.

Guzeldere, G. 1995. Problems of consciousness: A perspective on contemporary
issues, current debates. *Journal of Consciousness Studies* 2:112-43.
A summary of recent philosophical debates over consciousness, focusing on
the "what/where/who/why/how" questions, the explanatory gap, and the
stalemate between "essentialist" and "causal" intuitions.

Hannay, A. 1987. The claims of consciousness: A critical survey. *Inquiry*
30:395-434.

Hannay, A. 1990. _Human Consciousness_. Routledge.

- Hurley, S. 1998. *Consciousness in Action*. Harvard University Press.
- Jackendoff, R. 1987. *Consciousness and the Computational Mind*. MIT Press.
Separates computational mind from phenomenological mind, and studies the former, a third-person approach. The residue is the "Mind-Mind" problem. Consciousness supervenes on an intermediate level of representation. Elegant.
- Kirk, R. 1994. *Raw Feeling: A Philosophical Account of the Essence of Consciousness*. Oxford University Press.
Physicalism can explain consciousness in all its glory. Argues against zombies and inverted-spectrum scenarios, and suggests that the explanatory gap can be bridged by an account of directly-active information-processing.
- Levine, J. 1997. Recent work on consciousness. *American Philosophical Quarterly* 34:379-404.
- Lycan, W.G. 1987. *Consciousness*. MIT Press.
- Lycan, W.G. 1996. *Consciousness and Experience*. MIT Press.
- Metzinger, T. (ed) 1995. *Conscious Experience*. Ferdinand Schoningh.
An excellent collection of 20 philosophical papers on consciousness.
- Murata, J. 1997. Consciousness and the mind-body problem. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.
- Nelkin, N. 1996. *Consciousness and the Origins of Thought*. Cambridge University Press.
- O'Shaughnessy, B. 2000. *Consciousness and the World*. Oxford University Press.
- Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. MIT Press.
- Revonsuo, A. & Kamppinen, M. (eds) 1994. *Consciousness in Philosophy and Cognitive Neuroscience*. Lawrence Erlbaum.
- Sayre, K.M. 1969. *Consciousness: A Philosophic Study of Minds and Machines*. Random House.
- Seager, W.E. 1999. *Theories of Consciousness: An Introduction and Assessment*. Routledge.
- Searle, J.R. 1989. Consciousness, unconsciousness, and intentionality. *Philosophical Topics* 17:193-209.
Argues that the first-person view has been ignored too much in the philosophy of mind. Even unconscious states are only mental by virtue of their potential consciousness.
- Searle, J.R. 1992. *The Rediscovery of the Mind*. MIT Press.
On the centrality of consciousness to the mind. Consciousness is irreducible but biological. On the history of the field, the structure of consciousness, its role in constituting intentionality, and problems with computation.
- Searle, J.R. 1993. The problem of consciousness. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.
On the notion of consciousness, its relation to the brain, and some features that need to be explained: its subjectivity, unity, intentionality, center and periphery, Gestalt structure, aspect of familiarity, and so on.
- Sheets-Johnstone, M. 1998. Consciousness: A natural history. *Journal of Consciousness Studies* 5:260-94.

Siewert, C. 1998. *The Significance of Consciousness*. Princeton University Press.

Smith, D.W. 1992. Consciousness in action. *Synthese* 90:119-43.

Sprigge, T.L.S. 1982. The importance of subjectivity: An inaugural lecture. *Inquiry* 25:143-63.

Value is only found within streams of consciousness. Three ways of studying it: phenomenology, anthropology, and by relation to the physical. With an analysis of the "self-transcending" nature of conscious intentionality.

Strawson, G. 1994. *Mental Reality*. MIT Press.

Sturgeon, S. 2000. *Matters of Mind: Consciousness, Reason and Nature*. Routledge.

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press.

Tye, M. 2000. *Consciousness, Color, and Content*. MIT Press.

Velmans, M. 1996. *The Science of Consciousness: Psychological, Neuropsychological, and Clinical Reviews*. Routledge.

Verges, F.G. 1974. Jackson on incorrigibility. *Australasian Journal of Philosophy* 52:243-50.

Villaneuva, E. (ed) 1991. *Consciousness: Philosophical Issues*. Ridgeview. A collection of philosophical articles on consciousness.

Young, A.W. & Block, N. 1997. Consciousness. In (V. Bruce, ed) *Unsolved Mysteries of the Mind: Tutorial Essays in Cognition*. Taylor and Francis.

1.1b The Concept of Consciousness

Armstrong, D.M. 1979. Three types of consciousness. In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.

Armstrong, D.M. 1981. What is consciousness? In *The Nature of Mind*. Cornell University Press.

On minimal consciousness, perceptual consciousness, and introspective consciousness. Introspective consciousness seems so special because it gives inner awareness of self, and memory of other mental events.

Baruss, I. 1986. Meta-analysis of definitions of consciousness. *Imagination, Cognition, and Personality* 6:321-29.

Bisiach, E. 1988. The (haunted) brain and consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Distinguishes C1 (phenomenal experience) from C2 (access of parts of a system to other parts). C2 is can be scientifically studied, and has a graspable, if fragmented, causal role. C1 is mysterious and perhaps beyond science.

Burt, C. 1962. The concept of consciousness. *British Journal of Psychology* 53:229-42.

Cam, P. 1985. Phenomenology and speech dispositions. *Philosophical Studies* 47:357-68.

Reportability is not phenomenology, as blindsight has reportability but no phenomenology.

Chalmers, D.J. 1997. Availability: The cognitive basis of experience? In (N. Block, O. Flanagan, and G. Guzeldere, eds) *The Nature of Consciousness*. MIT Press.

Argues that the cognitive correlate of consciousness is direct availability for global control.

Church, J. 1998. Two sorts of consciousness? *Communication and Cognition* 31:51-71.

Finkelstein, D.H. 1999. On the distinction between conscious and unconscious states of mind. *American Philosophical Quarterly* 36:79-100.

Gennaro, R.J. 1995. Does mentality entail consciousness? *Philosophia* 24:331-58.

Girle, R.A. 1996. Shades of consciousness. *Minds and Machines* 6:143-57.

Kirk, R. 1992. Consciousness and concepts. *Aristotelian Society Supplement* 66:23-40.

Analyzes consciousness in terms of the "presence" of information to the main decision-making processes of a system. No great conceptual capacities required, no higher-order thoughts. With application to blindsight.

Lormand, E. 1995. What qualitative consciousness is like. Manuscript.

Lormand, E. 1996. Nonphenomenal consciousness. *Nous* 30:242-61.

Manson, N. 2000. State consciousness and creature consciousness: A real distinction. *Philosophical Psychology* 13:405-410.

Matthews, G. 1977. Consciousness and life. *Philosophy* 52:13-26.

McBride, R. 1999. Consciousness and the state/transitive/creature distinction. *Philosophical Psychology* 12:181-196.

Moody, T.C. 1986. Distinguishing consciousness. *Philosophy and Phenomenological Research* 47:289-95.

Separates consciousness from the mental -- functionalist accounts work for the latter but not the former. With remarks on Zen "pure consciousness".

Natsoulas, T. 1978. Consciousness. *American Psychologist* 33:906-14.

On the role of consciousness in psychology, and distinguishing various notions of consciousness: mutual knowledge, internal knowledge, awareness, direct awareness, personal unity, wakefulness, and double consciousness.

Natsoulas, T. 1983. A selective review of conceptions of consciousness with special reference to behavioristic contributions. *Cognition and Brain Theory* 6:417-47.

Ideas about consciousness from Locke, Brentano, Hebb, Dennett, Skinner, Sellars, Aristotle, Gibson. Theories: inner eye vs. verbal vs. outer eye.

Natsoulas, T. 1983. Concepts of consciousness. *Journal of Mind and Behavior* 4:195-232.

Natsoulas, T. 1991. The concept of consciousness(1): The interpersonal meaning. *Journal for the Theory of Social Behavior* 21:63-89.

Natsoulas, T. 1991. The concept of consciousness(2): The personal meaning. *Journal for the Theory of Social Behavior* 21:339-67.

Natsoulas, T. 1992. The concept of consciousness(3): The awareness meaning. *Journal for the Theory of Social Behavior* 2:199-25.

Natsoulas, T. 1993. Consciousness(4): Varieties of intrinsic theory. *Journal of Mind and Behavior* 14:107-32.

Natsoulas, T. 1994. The concept of consciousness(4): The reflective meaning. *Journal for the Theory of Social Behavior* 24:373-400.

Natsoulas, T. 1994. The concept of consciousness(5): The unitive meaning. *Journal for the Theory of Social Behavior* 24:401-24.

Natsoulas, T. 1995. Consciousness(3) and Gibson's concept of awareness. *Journal of Mind and Behavior* 3:305-28.

Natsoulas, T. 1996-1998. The case for intrinsic theory (parts 1-3). *Journal of Mind and Behavior* 17:267-85, 17:369-89, 19:1.

Natsoulas, T. 1997. Consciousness and self-awareness: Consciousness(1,2,3,4,5,6). *Journal of Mind and Behavior* 18:53-94.

Nelkin, N. 1987. What is it like to be a person? *Mind and Language* 21:220-41. Critiques three senses of consc: awareness, verbalization and phenomenology. Argues that none are sufficient for person-consciousness. Quite good.

Nelkin, N. 1993. What is consciousness? *Philosophy of Science* 60:419-34. On three senses of consciousness: phenomenality, intentionality, and introspectibility. Argues from empirical evidence (especially blindsight cases) that these three are all dissociable.

O'Shaughnessy, B. 1991. The anatomy of consciousness. In (E. Villanueva, ed) *_Consciousness_*. Ridgeview.

Place, U.T. 1992. Two concepts of consciousness: The biological/private and the linguistic/social. *Acta Analytica* 7:53-72.

Rosenthal, D.M. 1990. The independence of consciousness and sensory quality. In (E. Villanueva, ed) *_Consciousness_*. Ridgeview. Argues that consciousness and sensory quality are independent properties: there can be unconscious sensations. Consciousness is a relational property.

Rosenthal, D.M. 1994. State consciousness and transitive consciousness. *Consciousness and Cognition* 2:355-63.

Shanon, B. 1990. Consciousness. *Journal of Mind and Behavior* 11:137-51. On three kinds of consciousness -- sensed being, mental awareness, and reflection -- and their relationships.

Tye, M. 1996. The burning house. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh. Uses various puzzles cases to distinguish higher-order consciousness, discriminatory consciousness, responsive consciousness, and phenomenal consciousness.

1.2 Explaining Consciousness?

1.2a Subjectivity and Objectivity (Nagel) [see also 1.3a]

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 4:435-50. Reprinted in *_Mortal Questions_* (Cambridge University Press, 1979). Physicalist explanations leave out consciousness, i.e. what it is like to be an organism. Objective accounts omit points of view (could there be an objective phenomenology?). Physicalism may be true, but we can't see how.

Nagel, T. 1979. Subjective and objective. In *_Mortal Questions_*. Cambridge University Press.

Subjective and objective views clash e.g. on meaning of life, free will, personal identity, mind-body problem, ethics. How to reconcile: reduction, elimination, annexation? Maybe just let multiple viewpoints coexist.

Nagel, T. 1986. *_The View From Nowhere_*. Oxford University Press.

Seeing philosophy as a clash between the subjective and objective views of various phenomena (mental states, self, knowledge, freedom, value, ethics). Eliminating the subjective is impossible.

Akins, K. 1993. What is it like to be boring and myopic? In (B. Dahlbom, ed) *_Dennett and his Critics_*. Blackwell.

Gives a detailed account of perceptual processing in bats, and suggests that we can know what bat-experience is like: it's like nerd experience. But then is there an unexplained residue?

Akins, K. 1993. A bat without qualities? In (M. Davies and G. Humphreys, eds) *_Consciousness: Psychological and Philosophical Essays_*. Blackwell.

On what science tells us about the experience of bats, birds, and others. Why a movie of bat-experience isn't good enough -- because of the inseparability of intentionality and experience. Science can do OK.

Baker, L.R. 1998. The first-person perspective: A test for naturalism. *American Philosophical Quarterly* 35:327-348.

Biro, J.I. 1991. Consciousness and subjectivity. In (E. Villanueva, ed) *_Consciousness_*. Ridgeview.

No real problems are posed by subjectivity and points of view, no matter how they are construed (fixed, portable, tokens, types). It's either a confusion or a triviality about the logic of indexicality.

Biro, J.I. 1993. Consciousness and objectivity. In (M. Davies and G. Humphreys, eds) *_Consciousness: Psychological and Philosophical Essays_*. Blackwell.

Carruthers, P. 2000. Sympathy and subjectivity. *Australasian Journal of Philosophy* 77:465-482.

Davis, L. 1982. What is it like to be an agent? *Erkenntnis* 18:195-213.

On what is required for consciousness of agency (rather than qualia): belief, intention, and most importantly desire, enabling a capacity to care. A robot could have all this, and it would be like something to be it.

Flanagan, O.J. 1985. Consciousness, naturalism and Nagel. *Journal of Mind and Behavior* 6:373-90.

Naturalism can do autophenomenology just fine.

Foss, J.E. 1989. On the logic of what it is like to be a conscious subject. *Australasian Journal of Philosophy* 67:305-320.

A Super Neuroscientist will know how we describe and think about experience, so will know as much as a Super Sympathist. One doesn't have to imagine to know what it's like. With remarks on bat experience.

Foss, J.E. 1993. Subjectivity, objectivity, and Nagel on consciousness. *Dialogue* 32:725-36.

Nagel conflates metaphysical and epistemological versions of the subjective/objective distinction. Consciousness is metaphysically subjective, and science is epistemically objective, so there is incompatibility.

Francescotti, R.M. 1993. Subjective experience and points of view. *Journal of Philosophical Research* 18:25-36.

Being graspable from only one point of view does not define the class of

facts about conscious experience. Various ways of cashing this out fail.

Haksar, V. 1981. Nagel on subjective and objective. *Inquiry* 24:105-21.

The objective and subjective don't conflict, but complement each other.

Hanna, P. 1990. Must thinking bats be conscious? *Philosophical Investigations* 13:350-55.

Hiley, D.R. 1978. Materialism and the inner life. *Southern Journal of Philosophy* 16:61-70.

Nagel conflates questions about sensory qualities with those about a unique point of view. The truth of physicalism is irrelevant to uniqueness.

Hill, C.S. 1977. Of bats, brains, and minds. *Philosophy and Phenomenological Research* 38:100-106.

Kekes, J. 1977. Physicalism and subjectivity. *Philosophy and Phenomenological Research* 37:533-6.

The subjective/objective distinction is ill-drawn. Objective descriptions aren't species-independent, but in terms of the space-time causal network. Science can explain the experience this way, but not provide the experience.

Lewis, D. 1983. Postscript to "Mad pain and Martian pain". In *Philosophical Papers*, Vol. 1. Cambridge University Press.

Knowing what it's like consists in an ability, not possession of information.

Lycan, W.G. 1987. "Subjectivity". In *Consciousness*. MIT Press.

Various anti-Nagel points.

Lycan, W.G. 1990. What is the "subjectivity" of the mental? *Philosophical Perspectives*.

The subjectivity of the mental is no more special than usual propositional subjectivity. It can be handled by a self-scanner model of introspection.

Malcolm, N. 1988. Subjectivity. *Philosophy* 63:147-60.

A critique of Nagel's idea of a "point of view" that is occupied by a "subject". There aren't any peculiar facts about given viewpoints.

Maloney, J.C. 1986. About being a bat. *Australasian Journal of Philosophy* 64:26-49.

Mandik, P. 2001. Mental representation and the subjectivity of consciousness. *Philosophical Psychology* 14:179-202.

Mellor, D.H. 1993. Nothing like experience. *Proceedings of the Aristotelian Society* 63:1-16.

There are no facts about what an experience is like. Knowing what it's like is an ability to imagine, recognize, and recall; this explains ineffability, etc. With remarks on the experience of imagining an experience.

McClamrock, R. 1992. Irreducibility and subjectivity. *Philosophical Studies* 67:177-92.

Phenomenological properties cannot be picked out in physical or computational terms; argues against Lycan's criticism of Nagel. But all this is compatible with materialism. With comments on the phenomenological tradition.

McCulloch, G. 1988. What it is like. *Philosophical Quarterly* 38:1-19.

Criticizes absent/inverted qualia arguments for a special "what it is like", but argues that the possibility of "what it is like" differences relative to semantic states shows that something's not conveyed by functional accounts.

McMullen, C. 1985. 'Knowing what it's like' and the essential indexical.

Philosophical Studies 48:211-33.

The Nagel/Jackson argument is analogous to the Perry indexical argument, and can be treated the same way.

Mounce, H.O. 1992. On Nagel and consciousness. *Philosophical Investigations* 15:178-84.

Muscari, P. 1985. The subjective character of experience. *Journal of Mind and Behavior* 6:577-97.

Muscari, P. 1987. The status of humans in Nagel's phenomenology. *Philosophical Forum* 19:23-33.

Nagel's dilemma: separating feeling from process. Moral consequences?

Nelkin, N. 1987. What is it like to be a person? *Mind and Language* 2:220-41.

Nagel-consciousness exists, but isn't so important. It's essential for sensations, but not for thoughts. Beings without it could still be persons.

Nemirow, L. 1980. Review of Nagel's *Mortal Questions*. *Philosophical Review* 89:473-7.

Understanding does not consist only in facts; we can understand via sympathy.

Nemirow, L. 1990. Physicalism and the cognitive role of acquaintance. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Knowing what it's like is really knowing how to imagine. We should reduce Nagel's question to a question about possession of a certain ability.

Pugmire, D. 1989. Bat or batman. *Philosophy* 64:207-17.

Subjectivity is not something we have knowledge of, as we lack comparisons.

Rorty, R. 1993. Holism, intrinsicity, and the ambition of transcendence. In (B. Dahlbom, ed) *Dennett and His Critics*. Blackwell.

On the Nagel/Dennett debate: Nagel holds out for unexplained intrinsic properties once the relational is all accounted for; Dennett can renounce the transcendental ambition. Remarks on realism, holism, and metaphilosophy.

Rudd, A.J. 1999. What it's like and what's really wrong with physicalism: A Wittgensteinian perspective. *Journal of Consciousness Studies* 5:454-63.

Russow, L. 1982. It's not like that to be a bat. *Behaviorism* 10:55-63.

Divides Nagel's problem: qualitative differences, special access, mineness.

Simoni-Wastila, H. 2000. Particularity and consciousness: Wittgenstein and Nagel on privacy, beetles and bats. *Philosophy Today* 44:415-425.

Taliaferro, C. 1988. Nagel's vista or taking subjectivity seriously. *Southern Journal of Philosophy* 26:393-401.

Nagel's 'View from Nowhere' doesn't take subjectivity seriously enough.

Teller, P. 1992. Subjectivity and knowing what it's like. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Rebutting various intuitions for the non-physical nature of experience. The Nagel/Jackson argument commits an intensional fallacy; experiences are physical states known from a different perspective.

Tilghman, B.R. 1991. What is it like to be an aardvark? *Philosophy* 66:325-38.

A Wittgensteinian critique of Nagel. Nagel's question is confused: "what it's like" is a matter of behavior, sociality, etc, not inner experience.

van Gulick, R. 1985. Physicalism and the subjectivity of the mental. *Philosophical Topics* 13:51-70.

Reducing doesn't imply understanding. Two different kinds of reduction.

Wider, K. 1989. Overtones of solipsism in Nagel's 'What is it like to be a bat?' and 'The view from nowhere'. *Philosophy and Phenomenological Research* 49:481-99.

Nagel is an epistemological solipsist, whether he likes it or not.

Wright, E. 1996. What it isn't like. *American Philosophical Quarterly* 33:23-42.

1.2b The Explanatory Gap (Levine)

Beckermann, A. 2000. The perennial problem of the reductive explainability of phenomenal consciousness: C.D. Broad on the explanatory gap. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Bieri, P. 1995. Why is consciousness puzzling? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Reflections on the explanatory gap between physical processes and conscious experience. With remarks on different sorts of consciousness, and on why we need an intelligible necessary connection.

Block, N. & Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review*.

Chalmers, D.J. & Jackson, F. 2001. Conceptual analysis and reductive explanation. *Philosophical Review*.

Ellis, R.D. & Newton, N. 1998. Three paradoxes of phenomenal consciousness: Bridging the explanatory gap. *Journal of Consciousness Studies* 5:419-42.

Hardin, C.L. 1987. Qualia and materialism: Closing the explanatory gap. *Philosophy and Phenomenological Research* 48:281-98.

On the physiological bases of phenomenal states, particularly color.

Inverted spectrum isn't really coherent, as coolness/warmth would have to be inverted too. So the contingency of qualia is diminished.

Hardin, C.L. 1992. Physiology, phenomenology, and Spinoza's true colors. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Argues in detail that psychophysics can provide a structural map to close the explanatory gap. If there is an explanatory residue, perhaps panpsychism can help.

Harnad, S. 1994. Why and how we are not zombies. *Journal of Consciousness Studies* 1:164-67.

Kim, J. 1998. Reduction, reductive explanation, and "the explanatory gap". Manuscript.

Argues for a distinction between reduction and reductive explanation, and argues that reductive explanations generally involves conceptual connections via "functionalization". With comments on Block and Stalnaker.

Levin, J. 1991. Analytic functionalism and the reduction of phenomenal states. *Philosophical Studies* 61:211-38.

Contra Kripkean arguments, a good enough functional theory may help close the conceivability/explanatory gap between the physical and qualia. Contra Nagel/Jackson, such a theory could provide us with recognitional abilities.

Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.

How do we explain the apparent contingency of the qualia-matter reduction? Even if it's not metaphysically contingent, it's conceptually contingent, so there's a gap in any physical explanation of qualia. Excellent.

Levine, J. 1993. On leaving out what it's like. In (M. Davies and G. Humphreys, eds) *_Consciousness: Psychological and Philosophical Essays_*. Blackwell.

Physical accounts leave out qualia epistemologically but not metaphysically. So physicalism holds, but there is an explanatory gap. Discusses Kripke's and Jackson's arguments in detail; also explanation and content.

Papineau, D. 1998. Mind the gap. *Philosophical Perspectives* 12:373-89.

Price, M.C. 1996. Should we expect to feel as if we understand consciousness? *Journal of Consciousness Studies* 3:303-12.

Argues that the explanatory gap between brain and consciousness is just the same as that found with causal relations everywhere; it's just that we usually overlook it in the latter case.

Sturgeon, S. 1994. The epistemic basis of subjectivity. *Journal of Philosophy* 91:221-35.

Qualia can't be explained in more basic terms, as that sort of explanation works by accounting for a property's canonical evidence, but the canonical evidence for qualia are qualia themselves. But they still may be physical.

Tye, M. 1999. Phenomenal consciousness: The explanatory gap as a cognitive illusion. *Mind* 108:705-25.

1.2c 'Hard' and 'Easy' Problems (Chalmers) [see also 3.1]

Chalmers, D.J. 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:200-19. Also in (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press. Reprinted in Shear 1997.

Divides the problems of consciousness into easy and hard problems; the hard problem eludes reductive explanation as it isn't about explaining functions. Argues instead for a nonreductive theory with psychophysical laws.

Chalmers, D.J. 1995. The puzzle of conscious experience. *Scientific American* 273(6):80-86.

Like the JCS article, but shorter, more accessible, and with pretty pictures.

Chalmers, D.J. 1996. Can consciousness be reductively explained? In *_The Conscious Mind_*. Oxford University Press.

There is no a priori entailment from the physical to phenomenal facts (arguments from conceivability, epistemology, analysis), so reductive explanation fails. With a critique of existing empirical proposals.

Chalmers, D.J. 1997. Moving forward on the problem of consciousness. *Journal of Consciousness Studies* 4:3-46. Reprinted in Shear 1997.

A reply to the 25 "hard problem" articles in JCS.

Chalmers, D.J. 1998. The problems of consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Churchland, P.S. 1996. The hornswoggle problem. *Journal of Consciousness Studies* 3:402-8. Reprinted in Shear 1997,

Argues that the "hard problem" in effect invokes an argument from ignorance, and that there's no deep difference between consciousness and other domains.

Clark, T. 1995. Function and phenomenology: Closing the explanatory gap. *Journal of Consciousness Studies* 2:241-54. Reprinted in Shear 1997.

Argues contra Chalmers that experience is identical to certain functions, rather than emerging from them.

- Crick, F. and Koch, C. 1995. Why neuroscience may be able to explain consciousness. *Scientific American* 273(6):84-85. Reprinted in Shear 1997.
Divides the hard problem into three parts, and argues that neuroscience can make progress on at least one part (incommunicability); and maybe "meaning" holds the key to the rest.
- Dennett, D.C. 1996. Facing backwards on the problem of consciousness. *Journal of Consciousness Studies* 3:4-6. Reprinted in Shear 1997.
Argues contra Chalmers 1995 that functions are all we need to explain.
- Eilan, N. 2000. Primitive consciousness and the 'hard problem'. *Journal of Consciousness Studies* 7:28-39.
- Gray, J. 1998. Creeping up on the hard question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Hodgson, D. 1996. The easy problems ain't so easy. *Journal of Consciousness Studies* 3:69-75. Reprinted in Shear 1997.
Argues that consciousness plays a vital role in performing mental functions, so the easy problems won't be solved until the hard problem is solved.
- Horst, S. 1999. Evolutionary explanation and the hard problem of consciousness. *Journal of Consciousness Studies* 6:39-48.
- Ismael, J. 1999. Science and the phenomenal. *Philosophy of Science* 66:351-69.
- Lewis, H. 1998. Consciousness: Inexplicable - and useless too? *Journal of Consciousness Studies* 5:59-66.
- Libet, B. 1996. Solutions to the hard problem of consciousness. *Journal of Consciousness Studies* 3:33-35. Reprinted in Shear 1997.
Endorses the idea of consciousness as fundamental, but criticizes Chalmers' psychophysical laws. Advocates a theory with a "conscious mental field".
- Lowe, E.J. 1995. There are no easy problems of consciousness. *Journal of Consciousness Studies* 2:266-71. Reprinted in Shear 1997.
Argues that the "easy problems" -- reportability, attention, etc -- all involve concepts and therefore experience itself, for Kantian reasons, and therefore are not mechanistically explainable.
- Mills, E.O. 1996. Giving up on the hard problem of consciousness. *Journal of Consciousness Studies* 3:26-32. Reprinted in Shear 1997.
Argues that the truly hard problem is that of giving a constitutive account of consciousness, and Chalmers doesn't solve that (laws aren't good enough); in fact it's unsolvable.
- Mills, F.B. 1998. The easy and hard problems of consciousness: A Cartesian perspective. *Journal of Mind and Behavior* 19:119-40.
- O'Hara, K. & Scutt, T. 1996. There is no hard problem of consciousness. *Journal of Consciousness Studies* 3:290-302. Reprinted in Shear 1997.
Argue that we should work on the easy problems for now, as nobody has any good ideas about the hard problem; maybe it will gradually fade away.
- Robinson, W.S. 1996. The hardness of the hard problem. *Journal of Consciousness Studies* 3:14-25. Reprinted in Shear 1997.
- Shear, J. 1996. The hard problem: Closing the empirical gap. *Journal of Consciousness Studies* 3:54-68. Reprinted in Shear 1997.
On the epistemology of the hard problem. Argues that a scientific study of phenomenology is possible, drawing on work in developmental psychology and Eastern thought. "Pure consciousness" may be relevant to a resolution.

Shear, J. (ed) 1997. Explaining Consciousness: The Hard Problem. MIT Press.

A collection of essays consisting of Chalmers' keynote paper, 26 replies from many perspectives, and Chalmers' response to the replies.

Varela, F. 1995. Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies* 3:330-49. Reprinted in Shear 1997. Advocates a careful phenomenological study of consciousness in its own right, systematically linked with a neurophysiological investigation.

Velmans, M. 1995. The relation of consciousness to the material world. *Journal of Consciousness Studies* 2:255-65. Reprinted in Shear 1997. Agrees with Chalmers on nonreductionism, but disagrees on "awareness", organizational invariance, and thermostats. Advocates a kind of dual-aspect theory, where the physical world is present within consciousness,

1.2d Cognitive Closure (McGinn)

Davies, W.M. 1999. Sir William Mitchell and the "new mysterianism". *Australasian Journal of Philosophy* 77:253-73.

Garvey, J. 1997. What does McGinn think we cannot know? *Analysis* 57:196-201.

Hanson, P.P. 1993. McGinn's cognitive closure. *Dialogue* 32:579-85.

Kirk, R. 1991. Why shouldn't we be able to solve the mind-body problem? *Analysis* 51:17-23.

McGinn asks too much of a solution to the M-B problem. We might understand consciousness without understanding specific experiences; we could get at it by studying brain and consciousness not separately but simultaneously.

Krellenstein, M.F. 1995. Unsolvable problems, visual imagery, and explanatory satisfaction. *Journal of Mind and Behavior* 16:235-54.

Kukla, A. 1995. Mystery, mind, and materialism. *Philosophical Psychology* 8:255-64.

McDonough, R. 1992. The last stand of mechanism. *Journal of Speculative Philosophy* 6:206-25.

McGinn, C. 1989. Can we solve the mind-body problem? *Mind* 98:349-66. Reprinted in The Problem of Consciousness (Blackwell, 1991).

Argues that the mind-body problem might be solvable in principle, but beyond human capacities. Neither perception of the brain nor introspection of consciousness can uncover the property by which consciousness arises.

McGinn, C. 1991. The Problem of Consciousness: Essays Toward a Resolution. Blackwell.

A collection of articles on the problem of consciousness, advocating a view on which the phenomenon is natural but permanently mysterious to us.

McGinn, C. 1991. Consciousness and the natural order. In The Problem of Consciousness. Blackwell.

Argues that a naturalistic account of the intentionality of conscious states requires an account of their embodiment; and that embodiment may depend on the hidden structure of conscious states, not accessible to introspection.

McGinn, C. 1991. The hidden structure of consciousness. In The Problem of Consciousness. Blackwell.

Suggests that consciousness may have a hidden structure, analogous to the deep structure of language, that relates its surface properties to physical

properties. We may not be able to understand this hidden structure, however.

McGinn, C. 1993. *Problems in Philosophy*. Blackwell.

McGinn, C. 1995. Consciousness and space. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

McGinn, C. 1999. *The Mysterious Flame: Conscious Minds in a Material World*. Basic Books.

Sacks, M. 1994. Cognitive closure and the limits of understanding. *Ratio* 7:26-42

Whitely, C.H. 1990. McGinn on the mind-body problem. *Mind* 99:289.

1.2e Miscellaneous

Churchland, P.M. 1996. The rediscovery of light. *Journal of Philosophy* 93:211-28.

Parodies arguments by Searle, Jackson, and Chalmers for the irreducibility of consciousness with analogous arguments for the irreducibility of "luminescence". The consciousness arguments are no better.

Churchland, P.S. 1998. What should we expect from a theory of consciousness? In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Hardcastle, V.G. 1993. The naturalists versus the skeptics: The debate over a scientific understanding of consciousness. *Journal of Mind and Behavior* 14:27-50.

Argues that consciousness can be handled within a scientific framework. We can translate first-person accounts into third-person accounts. Replies to skeptical objections using analogies from elsewhere in science.

Hardcastle, V.G. 1996. The why of consciousness: A non-issue for materialists. *Journal of Consciousness Studies* 3:7-13.

A "committed materialist" will not see any explanatory gap, or any "brute fact". The entrenched differences lie in one's choice of initial framework.

Hesslow, G. 1996. Will neuroscience explain consciousness? *Journal of Theoretical Biology* 171:29-39.

Kirk, R. 1995. How is consciousness possible? In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Kurthen, M. 1995. On the prospects of a naturalistic theory of phenomenal consciousness. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Lockwood, M. 1998. The enigma of sentience. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Nida-Rumelin, M. 1997. Is the naturalization of qualitative experience possible or sensible? In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Taylor, J.G, 1998. Cortical activity and the explanatory gap. *Consciousness and Cognition* 7:109-48.

van Gulick, R. 1993. Understanding the phenomenal mind: Are we all just armadillos? In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Qualia pose no insurmountable problems for materialism: knowledge argument can be answered, explanatory gap can be closed, and absent qualia arguments beg the question. With speculations on their functional role.

van Gulick, R. 1995. What would count as explaining consciousness? In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Distinguishes six explananda, four explanatory restrictions, and four sorts of relations between them, making 96 possible problems. With a discussion of whether and how the central problems might be answered.

1.3 Materialism and Dualism

1.3a The Knowledge Argument (Jackson) [see also 1.2a]

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Knowing a completed neuroscience does not imply knowing about qualia. Mary, the colorblind neuroscientist, gains color vision and learns about red. So physicalism is false, as there are facts over and above the physical facts.

Jackson, F. 1986. What Mary didn't know. *Journal of Philosophy* 83:291-5.

Reply to Churchland 1985: Mary *learns*, Churchland misstates the argument.

Alter, T. 1995. Mary's new perspective. *Australasian Journal of Philosophy* 73:585-84.

Contra Pereboom 1994: The way a color sensation appears is a fact about it.

Alter, T. 1998. A limited defense of the knowledge argument. *Philosophical Studies* 90:35-56.

Bachrach, J.E. 1990. Qualia and theory reduction: A criticism of Paul Churchland. *Iyyun* 281-94.

Argues that Churchland's neuroscientific descriptions must leave at least some qualia behind: they might account for what we know (e.g. brain states) in qualia-knowledge, but can't handle distinctions in how we know.

Bigelow, J. & Pargetter, R. 1990. Acquaintance with qualia. *Theoria*.

Mary gains knowledge of old facts, in a new way: she gains a new mode of acquaintance with those facts. Analogies with indexical knowledge: her new knowledge eliminates no possible worlds.

Churchland, P.M. 1985. Reduction, qualia and the direct introspection of brain states. *Journal of Philosophy* 82:8-28. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).

Qualia can undergo a normal reduction to the neurophysiological. Jackson commits an intensional fallacy; in any case, perhaps Mary can understand red. When we apprehend qualia, we are directly introspecting our brain state.

Churchland, P.M. 1989. Knowing qualia: A reply to Jackson. In *_A Neurocomputational Perspective_*. MIT Press.

Rejoinder to Jackson 1986. The key lies in knowing-how vs. knowing-that.

Conee, E. 1985. Physicalism and phenomenal properties. *Philosophical Quarterly* 35:296-302.

Contra Lewis, Nemirow, and Horgan on the knowledge argument. But qualia may still be physical (though outside vocab of science) due to their causal role.

Conee, E. 1994. Phenomenal knowledge. *Australasian Journal of Philosophy*.

Mary knew all the facts about qualia beforehand, she just wasn't acquainted with them. Knowledge by acquaintance isn't factual knowledge.

Cummins, R. 1984. The mind of the matter: Comments on Paul Churchland.
Philosophy of Science Association 1984, 2:791-8.

Speculation on how consciousness might be left out by a physical account.

Dennett, D.C. 1991. "Epiphenomenal" qualia? In *_Consciousness Explained_*, pp. 398-406. Little-Brown.

Argues that most people don't really imagine Mary's situation. In fact, Mary would be able to identify blue objects from the way they make her react.

Furash, G. 1989. Frank Jackson's knowledge argument against materialism.
Dialogue 32:1-6.

Defends Jackson's argument against criticisms by Nemirow, Smith & Jones, Warner, Horgan, & Conee. The argument forces physicalism into a quandary: either deny qualia, or make the confused claim that qualia are physical.

Gertler, B. 1999. A defense of the knowledge argument. *Philosophical Studies* 93:317-336.

Graham, G. & Horgan, T. 2000. Mary Mary, quite contrary. *Philosophical Studies* 99:59-87.

Harman, G. 1993. Can science understand the mind? In (G. Harman, ed) *_Conceptions of the Human Mind: Essays on Honor of George A. Miller_*. Lawrence Erlbaum.

On Dilthey's "Verstehen", or "understanding from within". Mostly about meaning, but with application to the knowledge Mary gains.

Hershfield, J. 1998. Lycan on the subjectivity of the mental. *Philosophical Psychology* 11:229-38.

Horgan, T. 1984. Jackson on physical information and qualia. *Philosophical Quarterly* 34:147-83.

Mary didn't know all the physical facts: she knew all the explicitly physical information, but not all the ontologically physical information.

Jacquette, D. 1995. The blue banana trick: Dennett on Jackson's color scientist. *Theoria* 61:217-30.

Kelly, J.S. 1989. On neutralizing introspection: The data of sensuous awareness. *Southern Journal of Philosophy* 27:29-53.

Lahav, R. 1994. A new challenge for the physicalist: Phenomenal indistinguishability. *Philosophia* 24:77-103.

A new version of the knowledge argument: given all the physical facts, one can't know when two experiences are indistinguishable. This avoids various objections to the standard version.

Levin, J. 1986. Could love be like a heatwave?: Physicalism and the subjective character of experience. *Philosophical Studies* 49:245-61. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Contra Nagel/Jackson: Understand qualia through relational properties, and separate the mental concept from the recognitional capacity.

Lewis, D. 1990. What experience teaches. In (W. Lycan, ed) *_Mind and Cognition_*. Blackwell.

Against the hypothesis that phenomenology carries information. If it does, then qualia are epiphenomenal. Better to analyze the "new information" as acquiring an ability instead. In-depth and entertaining.

Loar, B. 1990. Phenomenal states. *Philosophical Perspectives* 4:81-108.

Phenomenal and functional concepts are distinct, but the relevant properties may be identical. We directly refer to phenomenal properties by recognition. Remarks on other minds, transparency, incorrigibility & more. A meaty paper.

Lycan, W.G. 1995. A limited defense of phenomenal information. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.
Gives nine arguments against the Lewis/Nemirow ability analysis, and proposes instead that the same fact is learned a new way, like water and H₂O facts.
This sort of phenomenal information is no danger to materialism.

Lycan, W.G. 1998. Phenomenal information again: It is both real and intrinsically perspectival. *Philosophical Psychology* 11:239-42.

McConnell, J. 1995. In defense of the knowledge argument. *Philosophical Topics* 22:157-187.
Defends against objections from Dennett, Churchland, etc. Horgan's objection (same fact different ways) has a certain force, but the argument can be reformulated to avoid them and imply property dualism. With remarks on Loar.

Nemirow, L. 1995. Understanding rules. *Journal of Philosophy* 92:28-43.

Newton, N. 1986. Churchland on direct introspection of brain states. *Analysis* 46:97-102.
Contra Churchland 1985: we couldn't introspect sensations as brain states, although we could interpret them as such.

Nida-Rumelin, M. 1995. What Mary couldn't know: Belief about phenomenal states. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.
Reconstructing the Mary case as a Marianna case, and introducing a distinction between phenomenal and nonphenomenal beliefs, to support the knowledge argument. A very nice and sophisticated paper.

Nida-Rumelin, M. 1998. On belief about experiences: An epistemological distinction applied to the knowledge argument against physicalism. *Philosophy and Phenomenological Research* 58:51-73.

Papineau, D. 1993. Physicalism, consciousness, and the antipathetic fallacy. *Australasian Journal of Philosophy* 71:169-83.
Mary goes from a third-person concept of experience to a first-person concept, but they co-refer; we can refer to an experience without having the experience. Physical and phenomenal properties are brutally identical.

Papineau, D. 1995. The antipathetic fallacy and the boundaries of consciousness. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Pereboom, D. 1994. Bats, brain scientists, and the limits of introspection. *Philosophy and Phenomenological Research* 54:315-29.
Mary learns an old fact under a new mode of presentation, and doesn't even learn a new fact about a mode of presentation. Her access to internal states is always mediated by representation, so we can always ascend to a new mode.

Raymont, P. 1995. Tye's criticism of the knowledge argument. *Dialogue* 34:713-26.

Raymont, P. 1999. The know-how response to Jackson's knowledge argument. *Journal of Philosophical Research* 24:113-26.

Robinson, D. 1993. Epiphenomenalism, laws, and properties. *Philosophical Studies* 69:1-34.
A thorough discussion of Lewis 1990. Phenomenal information implies epiphenomenalism, even at the intra-psychic level. Remarks on ineffability, and on whether properties should be individuated by nomic role or by essence.

Robinson, H. 1993. Dennett on the knowledge argument. *Analysis* 53:174-7.
Contra Dennett, Mary can't tell an object's color unless she already knows

about experience. The knowledge argument bears on thought, not just qualia.

Robinson, H. 1993. The anti-materialist strategy and the "knowledge argument". In (H. Robinson, ed) *_Objections to Physicalism_*. Oxford University Press.

Shoemaker, S. 1984. Churchland on reduction, qualia, and introspection. *Philosophy of Science Association* 1984, 2:799-809.

Introspection reveals functional properties, not physical, so qualia should be reduced to the functional, not to the physical.

Stemmer, N. 1989. Physicalism and the argument from knowledge. *Australasian Journal of Philosophy* 67:84-91.

Physicalism explains all the relevant evidence, hence all facts, and needn't admit mental entities; belief in mental entities is based on physical facts.

Thompson, E. 1992. Novel colors. *Philosophical Studies* 68:321-49.

Interesting remarks on what it would be for someone to see colors that we cannot, combining philosophical considerations with empirical findings about color space. Argues that science could tell us what such colors are like.

Tye, M. 2000. Knowing what it is like: The ability hypothesis and the knowledge argument. In *_Consciousness, Color, and Content_*. MIT Press.

Warner, R. 1986. A challenge to physicalism. *Australasian Journal of Philosophy* 64:249-65.

A Jackson-like argument that physical knowledge can't give you the knowledge of what pain feels like. With detailed consideration of objections and replies. Argues from limited incorrigibility to factualism about pains.

Watkins, M. 1989. The knowledge argument against the knowledge argument. *Analysis*, 49:158-60.

Epiphenomenalism => qualia don't cause beliefs => we don't know about qualia.

Zemach, E. 1990. Churchland, introspection, and dualism. *Philosophia* 20:3-13.

1.3b Zombies & Modal Arguments [see also 1.2c, 1.3a, 1.3c, 1.3d, 1.7e]

Balog, K. 1999. Conceivability, possibility, and the mind-body problem. *Philosophical Review* 108:497-528.

Bringsjord, S. 1999. The zombie attack on the computational conception of mind. *Philosophy and Phenomenological Research* 59:41-69.

Brueckner, A. 2001. Chalmers' conceivability argument for dualism. *Analysis*.

Byrne, A. 1999. Cosmic hermeneutics. *Philosophical Perspectives*.

Chalmers, D.J. 1996. Naturalistic dualism. In *_The Conscious Mind_*. Oxford University Press.

Argues from the lack of logical supervenience to the falsity of physicalism. A two-dimensional analysis shows that objections from a posteriori necessity fail. Argues for a naturalistic variety of property dualism.

Chalmers, D.J. 2002. Does conceivability entail possibility? In (T. Gendler & J. Hawthorne, eds) *_Imagination, Conceivability, and Possibility_*. Oxford University Press.

Cottrell, A. 1999. Sniffing the camembert: On the conceivability of zombies. *Journal of Consciousness Studies* 6:4-12.

Dennett, D.C. 1995. The unimagined preposterousness of zombies. *Journal of Consciousness Studies* 2:322-26.

Guzeldere, G. 1995. Varieties of zombiehood. *Journal of Consciousness Studies* 2:326-33.

Hill, C.S. 1997. Imaginability, conceivability, possibility, and the mind-body problem. *Philosophical Studies* 87:61-85.
Argues that the conceivability of zombies and the like can be explained away, in terms of the cognitive separability of perceptual imagination and sympathetic imagination of the same states.

Hill, C.S. 1998. Chalmers on the apriority of modal knowledge. *Analysis* 58:20-26.

Hill, C.S. & McLaughlin, B.P. 1998. There are fewer things in reality than are dreamt of in Chalmers' philosophy. *Philosophy and Phenomenological Research*.

Kirk, R. 1974. Sentience and behaviour. *Mind* 81:43-60.
Describing a situation where we would be justified in believing in zombies.
Argues that zombies are logically possible, which seems incompatible with most or all varieties of materialism.

Kirk, R. 1974. Zombies vs materialists. *Aristotelian Society Supplement* 48:135-52.
Materialism requires that physical states logically entail all non-relational states; but zombies are logically possible, so materialism fails. With a description of a zombie, and replies to a verificationist. All very true.

Kirk, R. 1977. Reply to Don Locke on zombies and materialism. *Mind* 86:262-4.
Reply to Locke 1976: materialism needs zombies to be logically impossible.

Kirk, R. 1999. Why there couldn't be zombies. *Proceedings of the Aristotelian Society, Supplementary Volume* 73:1-16.

Kraemer, E.R. 1980. Imitation-man and the 'new' epiphenomenalism. *Canadian Journal of Philosophy* 10:479-487.
If Campbell's imitation man is possible, then the causal relation between the physical and phenomenal is unreliable.

Latham, N. 1998. Chalmers on the addition of consciousness to the physical world. *Philosophical Studies*.

Levine, J. 1998. Conceivability and the metaphysics of mind. *Nous* 32:449-480.

Locke, D. 1976. Zombies, schizophrenics, and purely physical objects. *Mind* 83:97-99.
Contra Kirk: the logical possibility of zombies is compatible with empirical materialism. With some comments on Kirk's thought-experiment.

Marton, P. 1998. Zombies vs. materialists: The battle over conceivability. *Southwest Philosophy Review* 14:131-38.

Melnyk, A. 1998. Physicalism unfalsified: Chalmers' inconclusive argument for dualism. In (B. Loewer & C. Gillett, eds) *_Physicalism and its Discontents_*. Oxford University Press.

Moody, T. 1994. Conversations with zombies. *Journal of Consciousness Studies* 1:196-200.
Argues that behavioral differences in zombies would show up, in their discourse about consciousness.

Nagel, T. 1998. Conceiving the impossible and the mind-body problem. *Philosophy* 73:337-52.

Perkins, M. 1970. Matter, sensation, and understanding. *American Philosophical Quarterly* 8:1-12.

On the possibility of an Insentient Perceiver, who perceives the world without sensation. Sensation is inessential to perception and understanding, except understanding in the "whatlike" manner.

Perkins, M. 1971. Sentience. *Journal of Philosophy* 68:329-37.

Argues for the conceivability of insentient perception of colors (in "Insent", a kind of blindsighter or zombie), in order to argue for a realistic account of colors.

Perry, J. 2001. The zombie argument. In *_Knowledge, Possibility, and Consciousness_*. MIT Press.

Perry, J. 2001. The modal argument. In *_Knowledge, Possibility, and Consciousness_*. MIT Press.

Prudovsky, G. 1995. Arguments from conceivability. *Ratio* 8:63-69.

Robb, D. 1999. Conceivability and consciousness. *Philosophical Topics*.

Robinson, H. 1976. The mind-body problem in contemporary philosophy. *Zygon* 11:346-360.

A discussion of materialism and its difficulties. The conceivability of zombies poses special problems. Criticism of Smart's & Armstrong's analyses.

Squires, R. 1974. Zombies vs materialists II. *Aristotelian Society Supplement* 48:153-63.

Stalnaker, R. 2002. What is it like to be a zombie? In (T. Gendler & J. Hawthorne, eds) *_Imagination, Coceivability, and Possibility_*. Oxford University Press.

Stoljar, D. 2000. Physicalism and the necessary a posteriori. *Journal of Philosophy* 87:33-55.

Thomas, N. 1998. Zombie killer. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

van Gulick, R. 1999. Conceiving beyond our means: The limits of thought experiments. In (S. Hameroff, A. Kaszniak, & D. Chalmers, eds) *_Toward a Science of Consciousness III_*. MIT Press.

Yablo, S. 1998. Concepts and consciousness. *Philosophy and Phenomenological Research*.

Yablo, S. 1998. Textbook Kripkeanism and the open texture of language. *Philosophical Quarterly*.

1.3c Essentialist Arguments (Kripke) [see also 1.3b, 1.3d]

Kripke, S.A. 1971. Identity and necessity. In (M. Munitz, ed) *_Identity and Individuation_*.

An identity between mental and physical states can't be contingent, as it relates rigid designators. But nevertheless the co-occurrence of certain mental and physical states is contingent, so the identity theory is false.

Kripke, S.A. 1972. *_Naming and Necessity_*. Harvard University Press.

Both "pain" and "C-fibres firing" are rigid designators, so if they are identical, this must be necessary. But their co-occurrence is contingent, and this can't be explained away epistemically, so the identity theory fails.

- Barnette, R. 1977. Kripke's pains. *Southern Journal of Philosophy* 15.
Argues that pain and the associated epistemic situation are inequivalent.
Beliefs about pain are simply produced by mechanisms, and could come about without any sensation.
- Bayne, S.R. 19xx. Kripke's Cartesian argument. *Philosophia*.
Trying to turn Kripke's argument against him: it's possible that pains and C-fibre stimulations are identical, so it's necessary that they're identical.
- Bealer, G. 1994. Mental properties. *Journal of Philosophy* 91:185-208.
On four arguments against the identity theory: multiple-realizability, modal, knowledge, and certainty arguments. All face difficulties due to scientific essentialism, but the latter two can be reformulated to avoid them.
- Blumenfeld, J. 1975. Kripke's refutation of materialism. *Australasian Journal of Philosophy* 53:151-6.
Kripke's argument doesn't refute token identity. Pains can have other essential properties besides painfulness, so psychophysical token identities can be necessary.
- Boyd, R. 1980. Materialism without reductionism: What physicalism does not entail. In (N. Block, ed) *_Readings in the Philosophy of Psychology_*, vol 1. Harvard University Press.
Materialism doesn't need rigid identities, due to the compositional plasticity of mental states. So the possibility of disembodiment is compatible with materialism. The possibility of zombies is illusory.
- Carney, J. & von Bretzel, P. 1973. Modern materialism and essentialism. *Australasian Journal of Philosophy* 51:78-81.
A materialist must deny essentialism to meet Kripke's argument.
- Carney, J. 1975. Kripke and materialism. *Philosophical Studies* 27:279-282.
Comments on Feldman 1974: Feldman's view requires rejection of Kripke's views on necessity, or a problematic mixed view on rigid designators.
- Della Rocca, M. 1993. Kripke's essentialist arguments against the identity theory. *Philosophical Studies* 69:101-112.
Kripke's premise that pains are essentially mental either begs the question (by assuming pains don't have physical properties) or weakens the premise that physical events aren't essentially mental.
- Double, R. 1976. The inconclusiveness of Kripke's argument against the identity theory. *Auslegung* 3:156-65.
- Feldman, F. 1973. Kripke's argument against materialism. *Philosophical Studies* 24:416-19.
Painfulness need not be an essential feature of pains.
- Feldman, F. 1974. Kripke on the identity theory. *Journal of Philosophy* 71:665-76.
Kripke's arguments against person-body and mind-brain identity rely on the essentialness of aliveness to persons and painfulness to pains. There's no reason to grant this. If we do, rigidity is irrelevant to the argument.
- Feldman, F. 1980. Identity, necessity, and events. In (N. Block, ed) *_Readings in the Philosophy of Psychology_*, Vol. 1. Harvard University Press.
Defending a contingent event identity thesis against Kripke. Mental properties (which are distinct from physical properties) may not be essential properties of an event.
- Gjelsvik, O. 1988. A Kripkean objection to Kripke's arguments against the identity-theories. *Inquiry* 30:435-50.

Uses Kripke's 1979 direct-reference theory against him. When rigid designators don't have associated reference-fixing descriptions, we can't expect the "explaining away" strategy to work.

Hill, C.S. 1981. Why Cartesian intuitions are compatible with the identity thesis. *Philosophy and Phenomenological Research* 42:254-65.

The apparent contingency of identity is due to the fact that one can be aware of pain without being aware of C-fibers and vice versa, as well as to the fact that "C-fibers" may be picked out by a contingent description.

Holman, E. 1988. Qualia, Kripkean arguments, and subjectivity. *Philosophy Research Archives* 13:411-29.

Defending Kripkean arguments against various objections. Analysis in terms of manifest properties and their role in fixing reference to the subjective and objective.

Jackson, F. 1980. A note on physicalism and heat. *Australasian Journal of Philosophy* 58:26-34.

A Kripkean argument against non-analytic physicalism. Even if pain rigidly designates a brain state, the physicalist still has problems explaining the property of "pain-presents".

Jacquette, D. 1987. Kripke and the mind-body problem. *Dialectica* 41:293-300.

Kripke's argument doesn't refute contingent identity between minds and nonrigidly designated bodies, which is all materialism needs.

Leplin, J. 1979. Theoretical identification and the mind-body problem. *Philosophia* 8:673-88.

Some theoretical identifications are analogous to mental-physical identifications -- entities are introduced by properties considered essential within a theory, but this doesn't preclude identification.

Levin, M. 1975. Kripke's argument against the identity thesis. *Journal of Philosophy* 72:149-67.

The reference of "pain" is fixed not by essential features but by contingent topic-neutral descriptions; this is the real moral of Wittgenstein's private language argument. So Kripke's apparent contingency can be explained away?

Levin, M. 1995. Tortuous dualism. *Journal of Philosophy* 92:313-22.

Reply to Bealer 1994. Tries to clarify the dialectic, and argues that the materialist can explain "possibility" of straw thought as thought conjoined with mere appearance of straw.

Lycan, W.G. 1974. Kripke and the materialists. *Journal of Philosophy* 71:677-89.

Kripke equivocates on "pain-sensation": pains aren't the same as impressions of pain. Argues that imaginability arguments aren't decisive, and that functionalism may be less vulnerable than the identity theory.

Lycan, W.G. 1987. Functionalism and essence. In *_Consciousness_*. MIT Press.

Painfulness needn't be essential to pains: pains are events, not objects, and events don't have essences; and the reference of "pain" is fixed by topic-neutral descriptions. With remarks on pains vs. pain-sensations.

Maxwell, G. 1979. Rigid designators and mind-brain identity. *Minnesota Studies in the Philosophy of Science* 9.

"Brain state" reference is fixed by topic-neutral description; picks out pain but mightn't have, explaining the illusion of contingency. "Nonphysicalist materialism" results, with mind the essence of matter. An important paper.

McGinn, C. 1977. Anomalous monism and Kripke's Cartesian intuitions. *Analysis* 2:78-80. Reprinted in (N. Block, ed) *_Readings in the Philosophy of*

Psychology_ (MIT Press, 1980).

Token identity theories aren't vulnerable to Kripke's argument: it may be essential to this pain that it is a C-fibre firing, although not to pain as a type.

McMullen, C. 1984. An argument against the identity theory. Pacific Philosophical Quarterly 65:277-87.

We can explain away the apparent contingency of identity in terms of possible differences in *evidence* for the physical state. With a discussion on identities between something perceived and something described.

Mucciolo, L. 1975. On Kripke's argument against the identity thesis. Philosophia 5:499-506.

"Pain" need not be a rigid designator, but instead may pick out a state by its causal role. If it is a rigid designator, then the apparent contingency of identity comes from imagining something else filling the causal role.

Sher, G. 1977. Kripke, Cartesian intuitions, and materialism. Canadian Journal of Philosophy 7:227-38.

The reference of "C-fibre stimulation" might be fixed contingently, allowing the intuitive contingency of identity to be explained away.

Taylor, P. 1983. McGinn, token physicalism, and a rejoinder of Woodfield. Analysis 43:80-83.

Woodfield, A. 1978. Identity theories and the argument from epistemic counterparts. Analysis 38:140-3.

Contra McGinn 1977, the counterpart strategy fails as any pain that occurred here now would have been this pain. A counterpart strategy on brain states may work. With a reply by McGinn and a later rejoinder by Woodfield.

1.3d Arguments from Disembodiment [see also 1.3b, 1.3c]

Alston, W.P. & Smythe, T.W. 1994. Swinburne's argument for dualism. Faith and Philosophy 11:127-33.

Carrier, L. 1974. Definitions and disembodied minds. Personalist Forum 55:334-43.

Cole, D.J. & Foelber, F. 1984. Contingent materialism. Pacific Philosophical Quarterly 65:74-85.

Argues that materialism is only contingently true, as it's conceptually possible that we could become immaterial by gradual replacement.

Hart, W.D. 1988. _The Engines of the Soul_. Cambridge University Press.

Lewy, C. 1943. Is the notion of disembodied existence self-contradictory? Proceedings of the Aristotelian Society 43:59-78.

Long, D. 1977. Disembodied existence, physicalism, and the mind-body problem. Philosophical Studies.

Merricks, T. 1994. A new objection to a priori arguments for dualism. American Philosophical Quarterly 31:81-85.

Physicalism is compatible with the possibility of disembodiment: one can hold that mind and body are identical, that the body is physical, but that it is not essentially physical.

Odegard, D. 1970. Disembodied existence and central state materialism. Australasian Journal of Philosophy 48:256-60.

Pecnjak, D. 1995. Remarks on disembodied existence. Acta Analytica 10:209-13.

Shoemaker, S. 19xx On an argument for dualism. Reprinted in *_Identity, Cause, and Mind_*. Cambridge University Press, 1984.

Swinburne, R. 1997. The modal argument for substance dualism. In *_The Evolution of the Soul_* (revised edition). Oxford University Press.

Taliaferro, C. 1986. A modal argument for dualism. *Southern Journal of Philosophy* 24:95-108.

Taliaferro, C. 1997. Possibilities in the philosophy of mind. *Philosophy and Phenomenological Research* 57:127-37.

Tidman, P. 1994. Conceivability as a test for possibility. *American Philosophical Quarterly* 31:297-309.

Tye, M. 1983. On the possibility of disembodied existence. *Australasian Journal of Philosophy* 61:275-282.

There's no reason to believe that disembodied existence is possible: lack of logical contradiction doesn't imply possibility, conceivability is too weak a criterion, and it's not obvious that the situation is imaginable.

van Cleve, J. 1983. Conceivability and the Cartesian argument for dualism. *Pacific Philosophical Quarterly*.

Yablo, S. 1993. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53:1-42.

Zimmerman, D. 1991. Two Cartesian arguments for the simplicity of the soul. *American Philosophical Quarterly* 28:127-37.

1.3e Consciousness and Physicalism, Misc [see also 1.2, 1.3, 1.7b, 3.5b]

Fox, M. 1978. Beyond materialism. *Dialogue* 17:367-70.

Hill, C.S. 1991. *_Sensations: A Defense of Type Materialism_*. Cambridge University Press.

Defending type materialism, by way of criticism of dualism and functionalism. With treatments of introspection, sensory concepts, and other minds.

Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, & S. White, eds) *_Philosophy and the Cognitive Sciences_*. Holder-Pichler-Tempsky.

On why materialism requires conceptual analysis to locate mental properties in the natural world. Even a posteriori necessary connections have to be backed by a priori links. With remarks on supervenience. A nice paper.

Kirk, R. 1979. From physical explicability to full-blooded materialism. *Philosophical Quarterly* 29:229-37.

If every physical events has a physical explanation, and the mental is causally efficacious, then mental facts are strictly implied by physical facts. A nice argument.

Kirk, R. 1982. Physicalism, identity, and strict implication. *Ratio* 24:131-41.

Physicalism requires that all mental facts be strictly implied by the physical facts. Once this is recognized, questions about necessary or contingent identity are beside the point, and indeed identity is irrelevant.

Locke, D. 1971. Must a materialist pretend he's anaesthetized? *Philosophical Quarterly* 21:217-31.

Lund, D. 2000. Materialism and the subject of consciousness. *Idealistic Studies* 30:7-23.

Madell, G. 1988. *Mind and Materialism*. Edinburgh University Press.
On the problems posed for materialism by intentionality, autonomy, awareness, and indexicality. Tentatively advocates a Cartesian position.

Robinson, H. 1982. *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Robinson, H. (ed) 1993. *Objections to Physicalism*. Oxford University Press.
A collection of arguments against physicalism, mostly based on worries about consciousness and qualia.

Robinson, W.S. 1982. Sellarsian materialism. *Philosophy of Science* 49:212-27.

Seager, W.E. 1992. *Metaphysics of Consciousness*. Routledge and Kegan Paul.
Consciousness could be physical even if not explicable; but supervenience worries make it hard to see how it *could* be physical, though causal role suggests that it must be. We need a new conception. A stimulating book.

Sellars, W. 1981. Is consciousness physical? *Monist* 64:66-90.
On the place of "occurrent pink" and the "sensorium" in the physical world. It may turn out that the physics of the brain differs from other physics, in order to accommodate the causal role of sensations.

Smith, A.D. 1993. Non-reductive physicalism? In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.
A careful discussion of how to characterize physicalism, in terms of identity or supervenience, and argues that physicalism must reduce (bowdlerize) qualia to something they are not, as physicalism requires topic-neutral analyses.

1.3f Consciousness and Dualism [see also 1.2c, 1.3, 3.4d]

Collins, C. 1997. Searle on consciousness and dualism. *International Journal of Philosophical Studies* 5:15-33.

Double, R. 1983. Nagel's argument that mental properties are nonphysical. *Philosophy Research Archives* 9:217-22.

Eccles, J. 1987. Brain and mind: Two or one? In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Foster, J. 1989. A defense of dualism. In (J. Smythies & J. Beloff, eds) *The Case for Dualism*. University of Virginia Press.
Argues that all forms of materialism fail, and that dualism is the only option. Defends dualism against objections, and argues for interactionism over epiphenomenalism. A very clear and interesting paper.

Honderich, T. 1981. Psychophysical law-like connections and their problems. *Inquiry* 24:277-303.
Defending lawlike connections between physical states & conscious occurrents. Contra anomalous monism and identity theory for occurrents. But occurrents may not be causally efficacious. Comments by Wilson/Sprigge/Mackie/Stich.

Lahav, R. & Shanks, N. 1992. How to be a scientifically respectable 'property dualist'. *Journal of Mind and Behavior* 13:211-32.
Argues that the hypothesis of consciousness as an irreducible global property of the brain is compatible with what we know of both neuroscience and physics. With interesting remarks on quantum mechanics.

Latham, N. 2000. Chalmers on the addition of consciousness to the physical

world. Philosophical Studies 98:67-93.

McGinn, C. 1993. Consciousness and cosmology: Hyperdualism ventilated. In (M. Davies and G. Humphreys, eds) *_Consciousness: Psychological and Philosophical Essays_*. Blackwell.

A dialogue with a "hyperdualist". On the pros and cons of materialist vs. dualist ontology and cosmology. Dualism avoids the "magic" of emergence at the cost of an inflated and bizarre ontology.

O'Leary-Hawthorne, J. & McDonough, J.K. 1998. Numbers, minds, and bodies: A fresh look at mind-body dualism. *Philosophical Perspectives* 12:349-71.

Rosenberg, J.F. 1988. On not knowing what or who one is: Reflections on the intelligibility of dualism. *Topoi* 7:57-63.

Smook, R. 1988. Egoicity and twins. *Dialogue* 27:277-86.

Smythies, J.R. & Beloff, J. (eds) 1989. *_The Case for Dualism_*. University of Virginia Press.

Sprigge, T.L.S. 1994. Consciousness. *Synthese* 98:73-93.

On the non-physical nature of consciousness, and the threat of a merely contingent connection to behavior; suggests a denial of "Hume's principle". Perhaps consciousness is the noumenal essence of the physical. A nice paper.

Swinburne, R. 1986. *_The Evolution of the Soul_*. Oxford University Press.

Taliaferro, C. 1996. *_Consciousness and the Mind of God_*. Cambridge University Press.

von Wright, G.H. 1994. On mind and matter. *Journal of Theoretical Biology* 171:101-10.

1.3g Mind-Body Problem, General

Beck, L.W. 1940. The psychophysical as a pseudo-problem. *Journal of Philosophy* 37:561-71.

Butler, C.W. 1972. The mind-body problem: A nonmaterialistic identity thesis. *Idealistic Studies* 2:229-48.

Campbell, K.K. 1970. *_Body and Mind_*. Doubleday.

Carrier, M. & Mittelstrass, J. 1991. *_Mind, Brain, Behavior: The Mind-Body Problem and the Philosophy of Psychology_*. de Gruyter.

Cheng, C. (ed) 1975. *_Philosophical Aspects of the Mind-Body Problem_*. Hawaii University Press.

Cooper, W.E. 1977. Beyond materialism and back again. *Dialogue* 16:191-206.

Diaz, J. 2000. Mind-body unity, dual aspect, and the emergence of consciousness. *Philosophical Psychology* 13:393-403.

Feigl, H. 1934. Logical analysis of the psychophysical problem. *Philosophy of Science* 1:420-45.

Feinberg, T.E. 1997. The irreducible perspectives of consciousness. *Seminars in Neurology* 17:85-93.

Fodor, J.A. 1981. The mind-body problem. *Scientific American* 244:114-25.
An overview: behaviorism, identity theory, functionalism, etc.

Foss, J.E. 1987. Is the mind-body problem empirical? Canadian Journal of Philosophy 17:505-32.

Yes it is. Empirical evidence bears on materialism, property dualism, emergentism, functionalism, interactive dualism, idealism, etc.

Gomes, G. 1995. Self-awareness and the mind-brain problem. Philosophical Psychology 8:155-65.

Gunderson, K. 1970. Asymmetries and mind-body perplexities. Minnesota Studies in the Philosophy of Science 4:273-309.

The core of the mind-body problem is the first/third-person asymmetry. It's like a periscope trying to place itself between its crosshairs. But this doesn't imply any strong ontological consequences.

Honderich, T. 1989. Mind and Brain. Oxford University Press.

Honderich, T. 1995. Consciousness, neural functionalism, and real subjectivity. American Philosophical Quarterly 32:369-381.

Against "neural functionalism", and on how Searle's view reduces to either neural functionalism or property dualism.

Howard, D.J. 1986. The new mentalism. International Philosophical Quarterly 26:353-7.

Hutto, D.D. 1998. An ideal solution to the problems of consciousness. Journal of Consciousness Studies 5:328-43.

Kim, J. 1997. The mind-body problem: Taking stock after forty years. Philosophical Perspectives 11:185-207.

Kneale, M. 1950. What is the mind-body problem? Proceedings of the Aristotelian Society 50:105-22.

Kohler, W. 1960. The mind-body problem. In (S. Hook, ed) Dimensions of Mind. New York University Press.

Levin, M. 1979. Metaphysics and the Mind-Body Problem. Oxford University Press.

Lockwood, M. 1989. Mind, Brain, and the Quantum. Oxford University Press.
On the mind-body problem and physical theory. Against reductive physicalism; instead, experience is the intrinsic nature of the physical. Discusses the interpretation of quantum mechanics and much else. A very enjoyable book.

Lowe, E.J. 1996. Subjects of Experience. Cambridge University Press.

Lund, D.H. 1994. Perception, Mind, and Personal Identity: A Critique of Materialism. University Press of America.

Margolis, J. 1974. Reductionism and ontological aspects of consciousness. Journal for the Theory of Social Behavior 4:3-16.

Matson, W.I. 1966. Why isn't the mind-body problem ancient? In (P. Feyerabend & G. Maxwell, eds) Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl. University of Minnesota Press.

Matson, W.I. 1976. Sentience. University of California Press.

McMullen, T. 1997. Sperry on consciousness as an emergent causal agent. Australian Journal of Psychology 49:152-155.

Nagel, T. 1993. What is the mind-body problem? In Experimental and Theoretical Studies of Consciousness (Ciba Foundation Symposium 174). Wiley.
On ways in which we might locate consciousness within the natural world via

scientific study. Perhaps we need an wider conception of objective reality.

Nagel, T. 1994. Consciousness and objective reality. In (R. Warner & T. Szubka, eds) The Mind-Body Problem: A Guide to the Current Debate. Blackwell.

Nagel, T. 2001. The psychophysical nexus. In (P. Boghossian & C. Peacocke, eds) New Essays on the A Priori. Oxford University Press.

O'Shaughnessy, B. 1994. The mind-body problem. In (R. Warner & T. Szubka, eds) The Mind-Body Problem: A Guide to the Current Debate. Blackwell.

Pratt, J.B. 1936. The present status of the mind-body problem. *Philosophical Review* 65:144-56.

Reber, A. 1997. Caterpillars and consciousness. *Philosophical Psychology* 10:437-49.

Ripley, C. 1984. Sperry's concept of consciousness. *Inquiry* 27:399-423.
An in-depth analysis of Sperry's views on consciousness. Sperry is not a dualist; he simply believes in "structural causation" based on emergent properties. Thorough and interesting.

Robinson, W.S. 1988. Brains and People: An Essay on Mentality and its Causal Conditions. Temple University Press.

Rosenthal, D.M. (ed) 1971. Materialism and the Mind-Body Problem. Prentice-Hall.
A collection of essays from the 1960s on the identity theory, functionalism, eliminative materialism.

Senchuk, D.M. 1991. Consciousness naturalized: Supervenience without physical determinism. *American Philosophical Quarterly* 28:37-47.

Sellars, W. 1953. A semantical solution of the mind-body problem. *Methodos* 5:45-84. Reprinted in Pure Pragmatics and Possible Worlds. Ridgview, 1980.

Sellars, W. 1971. The double knowledge approach to the mind-body problem. *New Scholasticism* 45:269-89.

Silberstein, M. 1998. Emergence and the mind-body problem. *Journal of Consciousness Studies* 5:464-82.

Sperry, R.W. 1969. A modified concept of consciousness. *Psychological Review* 76:532-36.
Consciousness is an emergent property of brain dynamics that itself governs low-level flow of excitation. Midway between mentalism and materialism.

Sperry, R.W. 1980. Mind-brain interaction: Mentalism yes, dualism no. *Neuroscience* 5:195-206.
A summary of the position whereupon mental properties are emergent and have independent causal powers. With a contrast to Popper and Eccles' dualism.

Sperry, R.W. 1992. Turnabout on consciousness: A mentalist view. *Journal of Mind and Behavior* 13:259-80.
An account of the "new mentalist paradigm". Clarifies earlier work, comments on others' interpretations. The view is monist and functionalist, but consciousness is a distinct emergent quality with a "downward" causal role.

Strawson, G. 1994. The experiential and the non-experiential. In (R. Warner & T. Szubka, eds) The Mind-Body Problem: A Guide to the Current Debate. Blackwell.

Velmans, M. 1990. Consciousness, brain, and the physical world. *Philosophical Psychology* 3:77-99.

Wagner, S.J. 1994. Supervenience, recognition, and consciousness. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Warner, R. & Szubka, T. 1994. *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

A collection of 27 (mostly original) papers on the mind-body problem.

Weintraub, R. 1999. The spatiality of the mental and the mind-body problem. *Synthese* 117:409-17.

Wilson, D.L. 1976. On the nature of consciousness and of physical reality. *Perspectives in Biology and Medicine* 19:568-581.

Wisdom, J. 1957. Some main mind-body problems. *Proceedings of the Aristotelian Society* 60:187-210.

1.4 Specific Views on Consciousness [see also 1.2, 1.3, 1.5c]

1.4a Higher-Order Thought Approaches (Rosenthal, etc) [see also 6.2i]

Aquila, R. 1990. Consciousness as higher-order thoughts: Two objections. *American Philosophical Quarterly* 27:81-87.

Higher-order thought theories have two unacceptable consequences: one can notice one's hearing a sound without noticing one's consciousness of the sound; and one can unconsciously perceive one's surroundings as gloomy.

Byrne, A. 1997. Some like it HOT: consciousness and higher-order thoughts. *Philosophical Studies* 2:103-29.

Carruthers, P. 1989. Brute experience. *Journal of Philosophy* 258-69.

Argues for a distinction between conscious and non-conscious experiences, depending on whether one is conscious of the experience. Animal experiences are of the second kind, and therefore are not morally significant.

Carruthers, P. 1992. Consciousness and concepts. *Aristotelian Society Supplement* 66:41-59.

Advocates the "reflexive thinking" account of consciousness over Kirk's "presence" account. Availability for reflexive thinking is naturally necessary and sufficient for qualia. Interesting paper.

Carruthers, P. 1996. *Language, Thought, and Consciousness*. Cambridge University Press.

Carruthers, P. 1997. Fragmentary versus reflexive consciousness. *Mind and Language* 12:181-95.

Carruthers, P. 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.

Dretske, F. 1993. Conscious experience. *Mind* 102:263-283.

Against higher-order thought accounts: one can have a conscious experience without being aware that one is having it. With remarks on thing-awareness vs. fact-awareness and on "inner-sense" accounts.

Dretske, F. 1995. Are experiences conscious? In *Naturalizing the Mind*. MIT Press.

We're not conscious *of* our experience in general, but conscious *with* it.

Criticizes HOP theories (not conceptualized enough) and HOT theories (rules out animals; there's more in experience than thought).

Francescotti, R.M. 1995. Higher-order thoughts and conscious experience. *Philosophical Psychology*.

Argues that a higher-order thought is insufficient for consciousness, even with Rosenthal's constraint. A causal constraint is required, but the only strong enough such constraint doesn't work.

Gennaro, R.J. 1993. Brute experience and the higher-order thought theory of consciousness. *Philosophical Papers* 22:51-69.

Carruthers 1989 misanalyzes higher-order thought theory. There's no need for conscious HOTs, and not too much conceptual sophistication is required, so animals might have HOTs and therefore conscious pains.

Gennaro, R.J. 1996. *_Consciousness and Self-consciousness: A Defense of the Higher-Order Thought Theory of Consciousness_*. John Benjamins.

Guzeldere, G. 1996. Consciousness and the introspective link principle. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Guzeldere, G. 1995. Is consciousness the perception of what passes in one's own mind? In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

A critique of higher-order-perception theories of consciousness. They're either committed to a "representational divide" fallacy or collapse into higher-order-thought or first-order theories.

Jamieson, D. & Bekoff, M. 1992. Carruthers on nonconscious experience. *Analysis* 52:23-28.

Various points against Carruthers 1989. His examples of nonconscious experience are likely conscious, and the higher-order account is circular.

Jacob, P. 1996. State consciousness revisited. *Acta Analytica* 11:29-54.

Kobes, B.W. 1995. Telic higher-order thoughts and Moore's paradox. *Philosophical Perspectives* 9:291-312.

Levine, J. 1997. Are qualia just representations? (Critical notice of Tye.) *Mind and Language* 12:101-13.

Lycan, W.G. 1995. Consciousness as internal monitoring, I. *Philosophical Perspectives* 9:1-14.

Argues for a Lockean quasi-perceptual view of consciousness as internal monitoring via second-order states. Contra objections, e.g. Rey's point that it makes consciousness too prevalent -- consciousness isn't an on-off affair.

Lycan, W. 2001. A simple argument for a higher-order representation theory of consciousness. *Analysis* 61:3-4.

Mellor, D.H. 1978. Conscious belief. *Proceedings of the Aristotelian Society* 78:87-101.

Conscious belief (or assent) is believing that one believes. Addresses various objections, from self-deception and from consciousness of assent. Communication needs conscious belief, not just belief.

Mellor, D.H. 1980. Consciousness and degrees of belief. In (D.H. Mellor, ed) *_Prospects for Pragmatism_*. Cambridge University Press.

Natsoulas, T. 1992. Appendage theory -- pro and con. *Journal of Mind and Behavior* 13:371-96.

On various pros and cons of HOT theories, to do with reflexivity, objects of HOTs, introspection, and so on. With comparisons to "intrinsic" theories.

Natsoulas, T. 1992. Are all instances of phenomenal experience conscious in the sense of their being objects of inner (second-order) consciousness? *American Journal of Psychology* 105:605-12.

Natsoulas, T. 1993. What is wrong with the appendage theory of consciousness? *Philosophical Psychology* 6:137-54.

On three theories of our direct awareness of conscious states: mental-eye theories, self-intimational theories, and appendage theory. Appendage theory (i.e. HOT theory) is promising, but how does an HOT determine its object?

Natsoulas, T. 1993. The importance of being conscious. *Journal of Mind and Behavior* 14:317-40.

On the differences between first-order and second-order consciousness. Second-order consciousness is essential for communication and locomotion. With remarks on "nonconscious consciousness".

Nelkin, N. 1989. Unconscious sensations. *Philosophical Psychology* 2:129-41. Separates CN (phenomenological) from C1 (info-processing) and C2 (higher order beliefs). CN is a subset of CS (image-representation state). We are always C2 of CN states, but not of other CS states: unconscious sensations!

Nelkin, N. 1995. The dissociation of phenomenal states from apperception. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Argues that we need not be apperceptively aware of phenomenal states. Introspection leaves the matter open, but some empirical results (e.g. hue blindsight) and theoretical arguments support dissociability.

Ridge, M. 2001. Taking solipsism seriously: Nonhuman animals and meta-cognitive theories of consciousness. *Philosophical Studies* 103:315-340.

Rosenthal, D.M. 1986. Two concepts of consciousness. *Philosophical Studies* 49:329-59.

Consciousness should be construed neither as sensation nor intentionality, but as the existence of higher-order thoughts.

Rosenthal, D.M. 1997. A theory of consciousness. In (N. Block, O. Flanagan, and G. Guzeldere, eds) *_The Nature of Consciousness_*. MIT Press.

A conscious mental state is a state that is the subject of a higher-order thought. Consciousness is not essential to mentality, should be separated from sensory quality, and is not an intrinsic property of conscious states.

Rosenthal, D.M. 1990. Why are verbally expressed thoughts conscious? *Bielefeld Report*.

Because verbally expressing and reporting are easily and immediately connected for 1st-order thoughts. But not for 2nd-order thoughts. Hmmm.

Rosenthal, D.M. 1993. Thinking that one thinks. In (M. Davies and G. Humphreys, eds) *_Consciousness: Psychological and Philosophical Essays_*. Blackwell.

Conscious states are states that are the contents of higher-order thoughts. Express/report distinction: we report them, and express the HOT (which may be unconscious). Defense against dispositional and collapsing objections.

Rosenthal, D.M. 1993. Explaining consciousness. Manuscript.

Distinguishes the sense in which we are aware of conscious states; argues for the separation of consciousness and sensation; and outlines how higher-order thoughts might explain the what-it's-like of conscious states.

Rosenthal, D.M. 1993. Higher-order thoughts and the appendage theory of consciousness. *Philosophical Psychology* 6:155-66.

In response to Natsoulas, HOT theory needn't answer the general question of how intentional states determine their objects. With remarks on the

other alternatives and the dangers of self-intimation.

Rosenthal, D.M. 1995. Moore's paradox and consciousness. *Philosophical Perspectives* 9:313-33.

Rosenthal, D.M. 1997. Apperception, sensation, and dissociability. *Mind and Language* 2:206-23.

Rosenthal, D.M. 1998. Consciousness and metacognition. In (D. Sperber, ed) *_Metarepresentation_*. Oxford University Press.

Rosenthal, D.M. 2000. Consciousness, interpretation, and consciousness. *Protosociology* 14.

Rowlands, M. 2001. Consciousness and higher-order thoughts. *Mind and Language* 16:290-310.

Seager, W.E. 1994. Dretske on HOT theories of consciousness. *Analysis* 54:270-76.

Seager, W.E. 1999. HOT Theory: The mentalistic reduction of consciousness. In *_Theories of Consciousness: An Introduction and Assessment_*. Routledge.

Stamenov, M.I. 1997. Grammar, meaning, and consciousness: What sentence structure can tell us about the structure of consciousness. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

Stoerig, P. 1997. Phenomenal vision and apperception: Evidence from blindsight. *Mind and Language* 2:224-37.

Stone, J. 2001. What is it like to have an unconscious mental state? *Philosophical Studies* 104:179-202.

Thomasson, A. 2000. After Brentano: A one-level theory of consciousness. *European Journal of Philosophy* 8:190-210.

1.4b Dennett on Consciousness [see also 1.7c]

Akins, K. 1996. Lost the plot? Reconstructing Dennett's multiple drafts theory of consciousness. *Mind and Language* 11:1-43.

Akins, K, & Winger, S. 1996. Ships in the night: Churchland and Ramachandran on Dennett's theory of consciousness. In (K. Akins, ed) *_Perception_*. Oxford University Press.

Arbib, M.A. 1972. Consciousness: The secondary role of language. *Journal of Philosophy* 69.

Baker, L.R. 1995. Content meets consciousness. *Philosophical Topics* 22:1-22.

Block, N. 1995. What is Dennett's theory a theory of? *Philosophical Topics* 22:23-40.

Bricke, J. 1984. Dennett's eliminative arguments. *Philosophical Studies* 45:413-29.

Criticizing Dennett's accounts of pains, dreams, and images: in no case do his arguments earn their eliminative conclusions.

Bricke, J. 1985. Consciousness and Dennett's intentionalist net. *Philosophical Studies* 48:249-56.

Reportability is no good for capturing consciousness: it completely leaves out the qualitative content of conscious states.

Churchland, P.S. & Ramachandran, V.S. 1993. Filling in: Why Dennett is wrong. In (B. Dahlbom, ed) Dennett and His Critics. Blackwell.
Argues that Dennett's account of the blindspot and scotomas are wrong.
Neurophysiological data suggests that blind areas are represented explicitly; psychological data shows that it's not just "more of the same".

Clark, S.R.L. 1993. Minds, memes, and rhetoric. *Inquiry* 36:3-16.

Dennett, D.C. 1968. Content and Consciousness. Routledge.

Dennett, D.C. 1978. Reply to Arbib and Gunderson. In Brainstorms. MIT Press.
On various notions of awareness: contents of the speech center, contents directing behavior, and contents of attention. We have privileged access to one sort, but it is a different sort that plays the main role in control.

Dennett, D.C. 1978. Toward a cognitive theory of consciousness. *Minnesota Studies in the Philosophy of Science*, Vol. 9. Reprinted in Brainstorms (MIT Press, 1978).
Conscious contents are contents of a buffer memory to which a public report module has access. We only have conscious access to propositional judgments, not to underlying processes. With a cute functional diagram.

Dennett, D.C. 1979. On the absence of phenomenology. In (D. Gustafson & B. Tapscott, eds) Body, Mind, and Method. Kluwer.
There is no real phenomenology. There are only *judgments* about phenomenology, and nothing more is going on. We don't have privileged access to anything, except perhaps certain propositional episodes.

Dennett, D.C. 1982. How to study human consciousness empirically, or, nothing comes to mind. *Synthese* 53:159-80.
We can study consciousness by the method of heterophenomenology: studying the things we say about conscious states, which we can interpret as we interpret texts. Autophenomenology gives nothing extra. With comments by Rorty.

Dennett, D.C. 1988. The evolution of consciousness. Manuscript.
Consciousness is a virtual machine which evolved.

Dennett, D.C. 1991. Consciousness Explained. Little-Brown.
Argues against the "Cartesian Theatre", advocating a "multiple drafts" model of consciousness. Presents a detailed model of processes underlying verbal report, and argues that there is nothing else (e.g. qualia) to explain.

Dennett, D.C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15:183-201.
Using temporal anomalies in consciousness to support a "Multiple Drafts" theory of consciousness rather than a "Cartesian Theater". Contents of consciousness are wholly determined by effects on action/memory.

Dennett, D.C. 1993. *Precis of Consciousness Explained*. *Philosophy and Phenomenological Research* 53:889-931.
A discussion of Consciousness Explained, with comments by Tye, Jackson, Shoemaker, and Rosenthal, and a reply by Dennett.

Dennett, D.C. 1993. Living on the edge. *Inquiry* 36:135-59.
A reply to Clark, Fellows & O'Hear, Foster, Lockwood, Seager, Siewert, and Sprigge.

Dennett, D.C. 1993. Caveat emptor. *Consciousness and Cognition* 2:48-57.
A reply to Baars & McGovern, Mangan, Toribio.

Dennett, D.C. 1995. Is perception the "leading edge" of memory. In (A.

Spafadora, ed) Memory and Oblivion.

There is no "leading edge" of consciousness, separating perception and memory. With an analysis of metacontrast cases, etc.

Dennett, D.C. 1995. Get real. *Philosophical Topics* 22:505

Dennett, D.C. 1996. Seeing is believing -- or is it? In (K. Akins, ed) Perception. Oxford University Press.

Dennett, D. 2001. Are we explaining consciousness yet? *Cognition* 79:221-37.

Densmore, S. & Dennett, D.C. 1999. The virtues of virtual machines. *Philosophy and Phenomenological Research* 59:747-61.

Dretske, F. 1995. Differences that make no difference. *Philosophical Topics* 22:41-57.

Criticizes Dennett's first-person operationalism as Cartesian. There can be awareness without judgment -- e.g. non-epistemic perception. This comes from information or "micro-judgments", and is not conceptual.

Fellows, R. & O'Hear, A. 1993. Consciousness avoided. *Inquiry* 36: 73-91.

Foster, J. 1993. Dennett's rejection of dualism. *Inquiry* 36:17-31.

Gunderson, K. 1972. Content and Consciousness and the mind-body problem. *Journal of Philosophy* 69.

Hutto, D. 1995. Consciousness demystified: A Wittgensteinian critique of Dennett. *Monist* 78:464-79.

Jackson, F. 1993. Appendix A (for philosophers). *Philosophy and Phenomenological Research* 53:897-901.

Presses Dennett on the "truth-maker" question for materialists: what sort of physical fact makes it true that people are conscious?

Johnsen, B. 1997. Dennett on qualia and consciousness: A critique. *Canadian Journal of Philosophy* 27:47-82.

Kirk, R. 1993. "The best set of tools"? Dennett's metaphors and the mind-body problem. *Philosophical Quarterly* 43:335-43.

Joycean machines and multiple drafts turn out to shed no light on the question of what features make a conscious system conscious.

Lockwood, M. 1993. Dennett's mind. *Inquiry*.

Argues for a suitably sophisticated Cartesian Theatre, and against the identification of phenomenology with judgments.

Mangan, B. 1993. Dennett, consciousness, and the sorrows of functionalism. *Consciousness and Cognition* 2:1-17.

Marbach, E. 1988. How to study consciousness phenomenologically or quite a lot comes to mind. *Journal of the British Society for Phenomenology*, 19:252-268.

Marbach, E. 1994. Troubles with heterophenomenology. In (R. Casati, B. Smith, & S. White, eds) Philosophy and the Cognitive Sciences. Holder-Pichler-Tempsky.

McCauley, R.N. 1993. Why the blind can't lead the blind: Dennett on the blind spot, blindsight, and sensory qualia. *Consciousness and Cognition* 2:155-64.

Brings empirical evidence to bear against Dennett's "filling-in" account of the blindspot, and argues that blindsight and the blindspot aren't analogous.

McGinn, C. 1995. Consciousness evaded: Comments on Dennett. *Philosophical Perspectives* 9:241-49.

Nikolinakos, D. 2000. Dennett on qualia: The case of pain, smell and taste. *Philosophical Psychology* 13:505-522.

Puccetti, R. 1993. Dennett on the split-brain. *Psychology* 4(52).

Radner, D. 1994. Heterophenomenology: Learning about the birds and the bees. *Journal of Philosophy* 91:389-403.

Rey, G. 1995. Dennett's unrealistic psychology. *Philosophical Topics* 22:259-89.

Robinson, W.S. 1972. Dennett's analysis of awareness. *Philosophical Studies* 23:147-52.

Robinson, W.S. 1994. Orwell, Stalin, and determinate qualia. *Pacific Philosophical Quarterly* 75:151-64.

Dennett's Orwell/Stalin argument doesn't establish its conclusion, as "brain smear" is quite compatible with determinate qualia.

Rockwell, T. 1996. Awareness, mental phenomena, and consciousness: A synthesis of Dennett and Rosenthal. *Journal of Consciousness Studies* 3:463-76.

Rorty, R. 1972. Dennett on awareness. *Philosophical Studies* 23:153-62.

Rosenthal, D.M. 1993. Multiple drafts and higher-order thoughts. *Philosophy and Phenomenological Research* 53:911-18.

Rosenthal, D.M. 1994. First-person operationalism and mental taxonomy. *Philosophical Topics* 22:319-349.

Rosenthal, D.M. 1995. Multiple drafts and the facts of the matter. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

We can subtract first-person operationalism from Dennett's multiple-drafts account, giving a higher-order thought theory.

Rosenthal, D. 2000. Content, interpretation, and consciousness. *Protosociology* 14:67-84.

Seager, W.E. 1993. Verification, skepticism, and consciousness. *Inquiry*. An elucidation of Dennett's fundamental eliminativism about phenomenology, resting on verificationist arguments. Like many sceptical arguments, it ends up too powerful to be convincing.

Seager, W.E. 1999. Dennett, part I and II. In *Theories of Consciousness: An Introduction and Assessment*. Routledge.

Shoemaker, S. 1993. Lovely and suspect ideas. *Philosophy and Phenomenological Research* 53:903-908.

Siewert, C. 1993. What Dennett can't imagine and why. *Inquiry*. Argues that zombies are conceivable, via partial zombiehood in blindsight patients who respond unprompted. Dennett's arguments rely on a question-begging third-person absolutism.

Sprigge, T.L.S. 1993. Is Dennett a disillusioned zimbo? *Inquiry* 36:33-57.

Toribio, J. 1993. Why there still has to be a theory of consciousness. *Consciousness and Cognition* 2:28-47.

Criticizes behavioral, localist, and "intransitive" approaches to consciousness, and recommends a "transitive" metacognitive approach. But criticizes Dennett for not explaining subjective experience.

Tye, M. 1993. Reflections on Dennett and consciousness. *Philosophy and*

Phenomenological Research 53:891-6.

Argues that Dennett's verificationism begs the question, and that "seeming" cannot be identified with believing or judging.

Van Gulick, R. 1995. Dennett, drafts, and phenomenal realism. Philosophical Topics 22:443-55.

Wuketits, F. 1994. Consciousness explained -- or explained away? Acta Analytica 9:55-64.

1.4c Functionalism [see also 1.4a, 1.4b, 1.5c, 1.8, 3.4]

Levin, J. 1991. Analytic functionalism and the reduction of phenomenal states. Philosophical Studies 61:211-38.

Contra Kripkean arguments, a good enough functional theory may help close the conceivability/explanatory gap between the physical and qualia. Contra Nagel/Jackson, such a theory could provide us with recognitional abilities.

Mangan, B. 1998. Against functionalism: Consciousness as an information-bearing medium. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness II. MIT Press.

Marcel, A. 1988. Phenomenal experience and functionalism. In (A. Marcel & E. Bisiach, eds) Consciousness in Contemporary Science. Oxford University Press.

Myin, E. 1998. Holism, functionalism and visual awareness. Communication and Cognition, 31:3-19.

Perlis, D. 1995. Consciousness and complexity: The cognitive quest. Annals of Mathematics and Artificial Intelligence 14:309-21.

Shoemaker, S. 1993. Functionalism and consciousness. In Experimental and Theoretical Studies of Consciousness (Ciba Foundation Symposium 174). Wiley. Argues that introspective access is essential to many sorts of mental state, due to constitutive rationality requirements. Against a perceptual model of introspection; introspecting and introspected states are closer than that.

Schweizer, P. 1996. Physicalism, functionalism, and conscious thought. Minds and Machines 6:61-87.

van Gulick, R. 1988. A functionalist plea for self-consciousness. Philosophical Review 97:149-88.

How functionalism can handle consciousness: Self-consciousness is the possession of reflexive metapsychological information. This helps understand learning, representation, and belief. Phenomenal experience is still tricky.

1.4d Eliminativism [see also 1.7c, 2.1c, 3.5c]

Allport, A. 1988. What concept of consciousness? In (A. Marcel & E. Bisiach, eds) Consciousness in Contemporary Science. Oxford University Press.

Churchland, P.S. 1983. Consciousness: the transmutation of a concept. Pacific Philosophical Quarterly 64:80-95.

Experimental evidence against consciousness/introspection/transparency.

Dennett, D.C. 1976. Are dreams experiences? Philosophical Review 73:151-71. Reprinted in Brainstorms (MIT Press, 1978).

Argues that dreams might not be experienced, but rather be stored directly into memory (the "cassette-tape" theory of dreaming).

Dennett, D.C. 1979. The onus re experiences: A reply to Emmett. *Philosophical Studies* 35, 315-18.

Emmett, K. 1978. Oneiric experiences. *Philosophical Studies* 34:445-50.

Nikolinakos, D. 1994. General anesthesia, consciousness, and the skeptical challenge. *Journal of Philosophy* 2:88-104.

Consciousness is an indispensable concept in anesthesiology, and therefore (contra Churchland and Wilkes) is a scientifically legitimate kind. With empirical details and anesthesiological theory on levels of consciousness.

Rey, G. 1982. A reason for doubting the existence of consciousness. In (R. Davidson, S. Schwartz, & D. Shapiro, eds) *_Consciousness and Self-Regulation_*, Vol 3. Plenum Press.

One could make a machine, duplicating the usual abilities that go along with consciousness, but surely it wouldn't be conscious. So what are the conditions for consciousness? Maybe there are none.

Rey, G. 1986. A question about consciousness. In (H. Otto & J. Tuedio, eds) *_Perspectives on Mind_*. Kluwer.

A rerun of Rey 1982: An unconscious machine could duplicate all the obvious criteria for consciousness, so maybe even we aren't conscious. With remarks on the relation between our belief in consciousness and consciousness itself.

Rey, G. 1995. Toward a projectivist account of conscious experience. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

We "project" consciousness into ourselves and others. There are no explanation-transcendent phenomena for which there is non-question-begging evidence. With remarks on self-attribution and Wittgenstein.

Smith, D.W. 1986. In (H. Otto & J. Tuedio, eds) *_Perspectives on Mind_*. Kluwer.

Commentary on Rey 1986: we are directly aware of our consciousness. It's not a theoretical entity, but rather something to be explained.

Tienison, J.L. 1987. Brains are not conscious. *Philosophical Papers* 16:187-93.

A skeptical argument: single neurons are not conscious, and adding a neuron won't produce consciousness, so finite brains are not conscious.

Wilkes, K.V. 1984. Is consciousness important? *British Journal for the Philosophy of Science* 35:223-43.

No, and it's not very coherent either. It divides into awakesness, sensation, sensory experience, and propositional attitudes. Also a history of the term.

Wilkes, K.V. 1988. Yishi, Duh, Um and consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Wilkes, K.V. 1995. Losing consciousness. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Consciousness is not a tenable notion in either commonsense or scientific psychology; we should return instead to the "psuche".

Williams, D.C. 1934. Scientific method and the existence of consciousness. *Psychological Review* 41:461-79.

Williams, D.C. 1959. Mind as a matter of fact. *Review of Metaphysics* 13:205-25.

1.4e Epiphenomenalism [see also 1.3a, 1.3f, 1.6e, 3.6]

Bieri, P. 1992. Trying out epiphenomenalism. *Erkenntnis* 36:283-309.

Birnbacher, D. 1988. Epiphenomenalism as a solution to the ontological mind-body problem. *Ratio* 1:17-32.

Chalmers, D.J. 1996. The paradox of phenomenal judgment. In *The Conscious Mind*. Oxford University Press.

Considers major arguments against the causal or explanatory irrelevance of consciousness -- arguments from self-knowledge, memory, reference, etc -- and argues that none pose fatal flaws.

Creel, R. 1980. Radical epiphenomenalism: B.F. Skinner's account of private events. *Behaviorism* 8:31-53.

Dennett, D.C. 1991. "Epiphenomenal" qualia? In *Consciousness Explained*, pp. 398-406. Little-Brown.

Discusses two senses of "epiphenomenalism" -- "Huxley's" and "philosophical" varieties -- and argues that the philosophical sort is crazy. (N.B. Huxley actually subscribed to the "philosophical" variety.)

Double, R. 1979. Taylor's refutation of epiphenomenalism. *Journal of Critical Analysis* 8:23-28.

Flanagan, O. & Polger, G. 1998. Consciousness, adaptation, and epiphenomenalism. In (G. Mulhauser, ed) *Evolving Consciousness*.

Hodges, M. 1979. Meaning and the impotence hypothesis. *Review of Metaphysics* 32:515-29.

Horowitz, A. 1999. Is there a problem in physicalist epiphenomenalism? *Philosophy and Phenomenological Research* 59:421-34.

Huxley, T. 1874. On the hypothesis that animals are automata. *Fortnightly Review* 95:555-80. Reprinted in *Collected Essays*. London, 1893.

Hyslop, A. 1998. Methodological epiphenomenalism. *Australasian Journal of Philosophy* 78:61-70.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Uses the Mary thought-experiment to argue that qualia are epiphenomenal, and argues that epiphenomenalism is a tenable doctrine.

Kraemer, E.R. 1980. Imitation-man and the 'new' epiphenomenalism. *Canadian Journal of Philosophy* 10:479-487.

If Campbell's imitation man is possible, then the causal relation between the physical and phenomenal is unreliable.

Lachs, J. 1963. Epiphenomenalism and the notion of cause. *Journal of Philosophy* 60:141-45.

Lachs, J. 1963. The impotent mind. *Review of Metaphysics* 17:187-99.

Lachs, J. 1967. Angel, animal, machine: Models for man. *Southern Journal of Philosophy* 5:221-27.

Long, W. 1953. Comments on the alleged proof of epiphenomenalism. *British Journal for the Philosophy of Science* 3:355-58.

Pauen, M. 2000. Painless pain: Property dualism and the causal role of phenomenal consciousness. *American Philosophical Quarterly* 37:51-64.

Pecnjak, D. 1989. Epiphenomenalism and machines: A discussion of van Rooijen's critique of Popper. *British Journal for the Philosophy of Science* 40:404-8.

Popper, K.R. 1977. Some remarks on panpsychism and epiphenomenalism. *Dialectica* 31:177-86.

Puccetti, R. 1974. Physicalism and the evolution of consciousness. *Canadian Journal of Philosophy Supplement* 1:171-83.

Robinson, D. 1993. Epiphenomenalism, laws, and properties. *Philosophical Studies* 69:1-34.

Argues that phenomenal information implies epiphenomenalism, even at the intra-psychic level. With remarks on ineffability and on whether properties should be individuated by nomic role or by essence.

Robinson, W.S. 1982. Causation, sensation, and knowledge. *Mind* 91:524-40.
Argues that we can have non-inferential knowledge of sensations even if they don't make a causal difference to our beliefs, so epiphenomenalism is OK.
All theories have a similar problems.

Rudd, A. 2000. Phenomenal judgment and mental causation. *Journal of Consciousness Studies* 7:53-69.

van Rooijen, J. 1987. Interactionism and evolution: A critique of Popper. *British Journal for the Philosophy of Science* 38:87-92.

Wasserman, G.D. 1982. Materialism and mentality. *Review of Metaphysics* 35:715-30.

Wassermann, G. 1979. Reply to Popper's attack on epiphenomenalism. *Mind* 88:572-75.

Watkins, M. 1989. The knowledge argument against the knowledge argument. *Analysis*, 49:158-60.

Epiphenomenalism => qualia don't cause beliefs => we don't know about qualia.

Wisdom, J.O. 1954. Is epiphenomenalism refutable? *Proceedings of the 2nd International Congress of the International Union for the Philosophy of Science* 5:73-78.

Woodhouse, M. 1974. A new epiphenomenalism? *Australasian Journal of Philosophy* 52:163-69.

1.4f Interactionism [see also 1.3f, 3.3d]

Beloff, J. 1994. Minds and machines: A radical dualist perspective. *Journal of Consciousness Studies* 1:32-37.

Buncombe, M. 1995. *_The Substance of Consciousness: An Argument for Interactionism_*. Avebury.

Elitzur, A.C. 1989. Consciousness and the incompleteness of the physical explanation of behavior. *Journal of Mind and Behavior* 10:1-20.

Argues from the fact that we talk about consciousness to the conclusion that consciousness plays an active role, so physical laws must be incomplete.

Elitzur, A.C. 1990. Neither idealism nor materialism: A reply to Snyder. *Journal of Mind and Behavior*.

Elitzur, A.C. 1995. Consciousness can no longer be ignored. *Journal of Consciousness Studies* 2:353-58.

Foster, J. 1991. *_The Immaterial Self: A Defense of the Cartesian Dualist Conception of Mind_*. Routledge.

Hodgson, D. 1991. *_The Mind Matters: Consciousness and Choice in a Quantum*

World_. Oxford University Press.

Jackson, F. 1980. Interactionism revived? *Philosophy of Social Science* 10:316-23.

Libet, B. 1994. A testable theory of mind-brain interaction. *Journal of Consciousness Studies* 1:119-26.

Lindahl, B.I.B. & Arhem, P. 1996. Mind as a force field: Comments on a new interactionistic hypothesis. *Journal of Theoretical Biology* 171:111-22.

Popper, K. & Eccles, J. 1977. *The Self and Its Brain: An Argument for Interactionism*. Springer.

Popper, K. 1994. *Knowledge and the Body-Mind Problem: In Defence of Interaction*. Routledge.

Roelofs, H.D. 1955. A case for dualism and interactionism. *Philosophy and Phenomenological Research* 15:451-76.

Snyder, D. 1990. On Elitzur's discussion of the impact of consciousness on the physical world. *Journal of Mind and Behavior*.
Argues with Elitzur on quantum mechanics and consciousness. With response.

Warner, R. 1996. Facing ourselves: Incorrigeability and the mind-body problem. *Journal of Consciousness Studies* 3:217-30.

Wilson, D.L. 1999. Mind-brain interactionism and the violation of physical laws. *Journal of Consciousness Studies*.

1.4g Panpsychism [see also 1.4h]

Bjelland, A.G. 1982. Popper's critique of panpsychism and process proto-mentalism. *Modern Schoolman* 59:233-43.

Butler, C. 1978. Panpsychism: A restatement of the genetic argument. *Idealist Studies* 8:33-39.

Chalmers, D.J. 1996. Is experience ubiquitous? In *The Conscious Mind*. Oxford University Press.

There are no strong arguments against panpsychism, and good reason to take it seriously. Extrapolating the processing properties crucial for standard complex experience suggests that simple process may yield simple experience.

de Quincey, C. 1994. Consciousness all the way down? *Journal of Consciousness Studies* 1:217-29.

An analysis of a debate between Griffin and McGinn on panexperientialism, arguing for new forms of understanding.

Edwards, P. 1967. Panpsychism. In (P. Edwards, ed) *The Encyclopedia of Philosophy*, volume 5. Macmillan.

An excellent review article on panpsychism; highly recommended.

Farleigh, P. 1998. Whitehead's even more dangerous idea. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Ford, M.P. 1981. William James: Panpsychist and metaphysical realist. *Transactions of the Peirce Society* 17:158-70.

Griffin, D.R. 1997. Panexperiential physicalism and the mind-body problem. *Journal of Consciousness Studies* 4:248-68.

An interesting paper arguing for an experiential aspect in all matter,

explicating a Whiteheadian position.

Griffin, D.R. 1998. *_Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem_*. University of California Press.

Griffin, D.R. 1998. Pantemporalism and panexperientialism. In (P. Harris, ed) *_The Textures of Time_*. University of Michigan Press.

Hartshorne, C. 1978. Panpsychism: Mind as sole reality. *Ultim Real Mean* 1:115-29.

Hut, P. & Shepard, R. 1996. Turning the "hard problem" upside-down and sideways. *Journal of Consciousness Studies* 3:313-29.

Argues for a new fundamental feature ("X") which stands to consciousness as time stands to motion, thus making consciousness possible and ubiquitous.

Nagel, T. 1979. Panpsychism. In *_Mortal Questions_*. Cambridge University Press.

Material composition, nonreductionism, realism, non-emergence => panpsychism.

Popper, K.R. 1977. Some remarks on panpsychism and epiphenomenalism. *Dialectica* 31:177-86.

Rensch, B. 1977. Argument for panpsychist identism. In (J. Cobb & D. Griffin, eds) *_Mind in Nature_*. University Press of America.

Robinson, E.A. 1949. Animism as a world hypothesis. *Philosophical Review* 58:53-63.

Rosenberg, G.H. 1996. Rethinking nature: A hard problem within the hard problem. *Journal of Consciousness Studies* 3:76-88.

On why consciousness extends beyond the cognitive. Argues that fundamental laws for consciousness must connect at a basic level, and argues that panpsychism is not as implausible as often thought.

Seager, W. 1995. Consciousness, information, and panpsychism. *Journal of Consciousness Studies* 2:272-88. Reprinted in (J. Shear, ed) *_Explaining Consciousness: The Hard Problem_* (MIT Press, 1999).

Examines a position on which experience is fundamental to the world, and suggests that this ought to lead to panpsychism. With some connections to information and quantum mechanics.

Sellars, R.W. 1960. Panpsychism or evolutionary materialism. *Philosophy of Science* 27:329-49.

Shepherd, J.J. 1974. Panpsychism and parsimony. *Process Studies* 4:3-10.

Sprigge, T.L.S. 1983. The vindication of panpsychism. In *_The Vindication of Absolute Idealism_*. Edinburgh University Press.

van Cleve, J. 1990. Mind -- dust or magic? Panpsychism versus emergence. *Philosophical Perspectives* 4:215-226.

On Nagel 1979: emergence is more plausible than panpsychism. A construal of emergence as nomological supervenience without logical supervenience.

Wright, S. 1953. Gene and organism. *American Naturalist*.

Wright, S. 1977. Panpsychism and science. In (J. Cobb & D. Griffin, eds) *_Mind in Nature_*. University Press of America.

1.4h Intrinsic Monism (Russell, etc) [see also 1.4h]

Blackburn, S. 1992. Filling in space. *Analysis* 52:62-3.

Physics is dispositional, but if there are only bare dispositions, then the world has no nature of its own. And if there are categorical grounds, we have no idea what they could be, except maybe subjective qualia.

Chalmers, D.J. 1996. The metaphysics of information. In *The Conscious Mind*, pp. 301-8 (see also pp. 153-55). Oxford University Press.

An "it from bit" view fits the Russellian metaphysics (described earlier): physics is info from the outside, (proto)experience is info from the inside. The problem is constituting macrophenomenal from microphenomenal; some ideas.

Demopolous, W. & Friedman, M. 1989. The concept of structure in Russell's *The Analysis of Matter*. In (C. Savage & C. Anderson, eds) *Rereading Russell: Essays in Bertrand Russell's Metaphysics and Epistemology*. University of Minnesota Press.

A nice account of Russell's (and Schlick's and Carnap's) structuralism and Newman's objection, with analysis. (N.B. no philosophy of mind.)

Feigl, H. 1958. The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science* 2:370-497. Reprinted (with a postscript) as *The 'Mental' and the 'Physical'*. University of Minnesota Press, 1967.

A long and very interesting essay on the mind-body problem. Ultimately advocates a "structural" view of the physical and identifies experience with the underlying reality, at least for some neurophysiological states.

Feigl, H. 1960. The mind-body problem: Not a pseudo-problem. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

Feigl, H. 1971. Some crucial issues of mind-body monism. *Synthese* 22:295-312.

Feigl, H. 1975. Russell and Schlick: A remarkable agreement on a monistic solution of the mind-body problem. *Erkenntnis* 9:11-34.

Argues that Russell's and Schlick's views on the structural nature of physics and the possible identification of the "content" with experience are quite close to each other. With interesting historical remarks.

Feser, E. 1998. Can phenomenal qualities exist unperceived? *Journal of Consciousness Studies* 4:405-14.

Foster, J. 1982. *The Case for Idealism*. Routledge.

Foster, J. 1991. Lockwood's hypothesis. In *The Immaterial Self*. Oxford University Press.

Argues that the Russellian view is implausible, as the structure of the underlying physical processes does not correspond to the structure and quality of consciousness.

Lockwood, M. 1981. What was Russell's neutral monism? *Midwest Studies in Philosophy* 6:143-58.

Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Oxford University Press. On the mind-body problem and physical theory. Against reductive physicalism; instead, experience is the intrinsic nature of the physical. Explores some potential links with quantum mechanics.

Lockwood, M. 1993. The grain problem. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

On the "grain problem" for the intrinsic-nature view: how do lots of microphysical qualities add up into a smooth experience? Appeals to quantum mechanics and a preferred set of observables.

Lockwood, M. 1998. Unsensed phenomenal qualities: A defence. *Journal of Consciousness Studies* 4:415-18.

Maxwell, G. 1971. Structural realism and the meaning of theoretical terms. Minnesota Studies in the Philosophy of Science.

Maxwell, G. 1978. Unity of consciousness and mind-brain identity. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Maxwell, G. 1979. Rigid designators and mind-brain identity. Minnesota Studies in the Philosophy of Science 9.

"Brain state" reference is fixed by topic-neutral description; picks out pain but mightn't have, explaining the illusion of contingency. "Nonphysicalist materialism" results, with mind the essence of matter. An important paper.

Newman, M.H.A. 1928. Mr. Russell's causal theory of perception. *Mind*. Argues against Russell's structuralism: any collection can be arranged to have a given structure, under some relation, so if physics tells us only about structure, it tells us at most the cardinality of the world.

Robinson, H. 1982. Matter: Turning the tables. In *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Rosenberg, G.H. 1997. *A Place for Consciousness: Probing the Deep Structure of the Natural World*. Dissertation, Indiana University.

Russell, B. 1927. *The Analysis of Matter*. London: Kegan Paul. Argues that physics characterizes the external world only structurally, and leaves intrinsic qualities unspecified. Only experience acquaints us with anything intrinsic. Perhaps the intrinsic nature of physics is experiential?

Schlick, M. 1925. *General Theory of Knowledge*.

Stoljar, D. 1997. Neutral monism. Manuscript.

Stoljar, D. 2001. Two conceptions of the physical. *Philosophy and Phenomenological Research* 62:253-81.

Stubenber, L. 1996. The place of qualia in the world of science. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Stubenber, L. 1997. Austria vs. Australia: Two versions of the identity theory. In (K. Lehrer & J. Marek, eds) *Austrian Philosophy, Past and Present*. Kluwer.

Stubenber, L. 1998. *Consciousness and Qualia*. John Benjamins.

1.5 Consciousness and Content

1.5a Consciousness and Intentionality (Searle, etc)

Cole, D. 1994. Thought and qualia. *Minds and Machines* 4:283-302.

Crane, T. 1998. Intentionality as the mark of the mental. *Philosophy*.

Davies, M. 1995. Consciousness and the varieties of aboutness. In (C. Macdonald, ed) *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford University Press.

Dunlop, C. 2000. Searle's unconscious mind. *Philosophical Psychology* 13:123-148.

Fodor, J. & Lepore, E. 1994. What is the Connection Principle? *Philosophy and Phenomenological Research* 54:837-45.

Searle's formulation of the connection principle is unclear, and there is no formulation is both plausible and interesting.

Georgalis, N. 1996. Awareness, understanding, and functionalism. *Erkenntnis* 44:225-56.

Gillett, E. 1996. Searle and the "deep unconscious". *Philosophy, Psychiatry, and Psychology* 3:191-200.

Gunderson, K. 1990. Consciousness and intentionality: Robots with and without the right stuff. In (C.A. Anderson & J. Owens, eds) *Propositional Attitudes: The Role of Content in Language, Logic, and Mind*. CSLI.

Marbach, E. 1993. *Mental Representation and Consciousness: Toward a Phenomenological Theory of Representation and Reference*. Kluwer.

Meijers, A. 2000. Mental causation and Searle's impossible conception of unconscious intentionality. *International Journal of Philosophical Studies* 8:155-170.

McLoughlin, J. 1999. Unwittingly recapitulating Freud: Searle's concept of a vocabulary of the unconscious. *Ratio* 12:34-53.

Natsoulas, T. 1992. Intentionality, consciousness, and subjectivity. *Journal of Mind and Behavior* 13:281-308.

Nelkin, N. 1989. Propositional attitudes and consciousness. *Philosophy and Phenomenological Research* 49:413-30.

About conscious beliefs. We are not "conscious of" beliefs, merely "conscious that" -- i.e. belief is not phenomenological.

Nelkin, N. 1993. The connection between intentionality and consciousness. In (M. Davies and G. Humphreys, eds) *Consciousness: Psychological and Philosophical Essays*. Blackwell.

Against Searle: some intentional states aren't even potentially conscious (blindsight, etc) and intentional content doesn't require a particular phenomenal feel. So there's no essential link. With remarks on McGinn.

Schweizer, P. 1994. Intentionality, qualia, and mind/brain identity. *Minds and Machines* 4:259-82.

Seager, W.E. 1999. Conscious intentionality and the anti-Cartesian catastrophe. In *Theories of Consciousness: An Introduction and Assessment*. Routledge.

Searle, J.R. 1984. Intentionality and its place in nature. *Synthese*.
(Subjective) intentionality sure is real. It causes and is caused.

Searle, J.R. 1990. Consciousness, explanatory inversion and cognitive science. *Behavioral and Brain Sciences* 13:585-642.

Advocates a "connection principle": intentional states must be potentially conscious. If not, they're brutally neurophysiological. So cog-sci talk of "intentional" cognitive mechanisms below the conscious level isn't justified.

Searle J.R. 1994. The connection principle and the ontology of the unconscious: A reply to Fodor and Lepore. *Philosophy and Phenomenological Research* 54:847-55.

Clarifying the connection principle -- it's necessary in order to see how certain nonconscious neural states qualify as unconscious mental states.

van Gulick, R. 1988. Consciousness, intrinsic intentionality, and self-understanding machines. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

van Gulick, R. 1995. Why the connection argument doesn't work. *Philosophy and Phenomenological Research* 55:201-7.

van Gulick, R. 1995. How should we understand the relation between intentionality and phenomenal consciousness. *Philosophical Perspectives* 9:271-89.

Worley, S. 1997. Belief and consciousness. *Philosophical Psychology* 10:41-55.
Argues that belief requires consciousness, as we can't make sense of the personal/subpersonal content distinction without appealing to consciousness.

1.5b The Content of Experience

Baldwin, T. 1992. The projective theory of sensory content. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Bermudez, J.L. 1994. Peacocke's argument against the autonomy of nonconceptual representational content. *Mind and Language* 9:402-18.

Bermudez, J. 2000. Naturalized sense data. *Philosophy and Phenomenological Research* 61:353-374.

Bermudez, J.L. 1995. Nonconceptual content: From perceptual experience to subpersonal computational states. *Mind and Language* 10:333-69.

Berger, G. 1987. On the structure of visual sentience. *Synthese* 71:355-70.

Bilgrami, A. 1994. On McDowell on the content of perceptual experience. *Philosophical Quarterly* 44:206-13.

Brewer, B. 1999. *Perception and Reason*. Oxford University Press.

Butchvarov, P. 1980. Adverbial theories of consciousness. *Midwest Studies in Philosophy* 5:261-80.

Church, J. 2000. 'Seeing as' and the double bind of consciousness. *Journal of Consciousness Studies* 7:99-112.

Clark, R. 1973. Sensuous judgments. *Nous* 7:45-56.

Clark, R. 1976. The sensuous content of perception. In (H. Castaneda, ed) *Action, Knowledge, and Reality*. Bobbs-Merrill.

Clark, R. 1981. Sensing, perceiving, thinking. *Grazer Philosophische Studien* 12:273-95.

Crane, T. (ed) 1992. *The Contents of Experience: Essays on Perception*. Cambridge University Press.

Crane, T. 1992. The nonconceptual content of experience. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Davies, W.M. 1996. *Experience and Content: Consequences of a Continuum Theory*. Avebury.

DeBellis, M. 1991. The representational content of musical experience. *Philosophy and Phenomenological Research* 51:303-24.
Contra Peacocke, we don't need sensational properties to understand the content of musical experience. Fine-grained representational properties can do the job, with the help of some Schenkerian analysis.

Gunther, Y.H. 1995. Perceptual content and the subpersonal. *Conference*

6:31-45.

Gunther, Y.H. 2001. Content, illusion, partition. *Philosophical Studies* 102:185-202.

Hamlyn, D.W. 1994. Perception, sensation, and non-conceptual content. *Philosophical Quarterly* 44:139-53.

Jackson, F. 1976. The existence of mental objects. *American Philosophical Quarterly* 13:33-40.

Kelly, S.D. 2001. The non-conceptual content of perceptual experience: Situation dependence and fineness of grain. *Philosophy and Phenomenological Research* 62:601-608.

Kraut, R. 1982. Sensory states and sensory objects. *Nous* 16:277-93.

Langsam, H. 2000. Experiences, thoughts, and qualia. *Philosophical Studies* 99:269-295.

Lowe, E.J. 1992. Experience and its objects. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Martin, M.G.F. 1992. Perception, concepts, and memory *Philosophical Review* 101:745-63.

McDowell, J. 1994. The content of perceptual experience. *Philosophical Quarterly* 44:190-205.

McFarland, D. 1998. Crane on concepts and experiential content. *Analysis* 58:54-58.

Millar, A. 1991. Concepts, experience, and inference. *Mind* 100:495-505.

Natsoulas, T. 1983. What are the objects of perceptual consciousness? *American Journal of Psychology* 96:435-67.

Natsoulas, T. 1994. On the distinction between the object and the content of consciousness. *Journal of Mind and Behavior* 15:239-64.

Noe, A. 1999. Thought and experience. *American Philosophical Quarterly* 36:257-65.

Peacocke, C. 1983. *Sense and Content: Experience, Thought, and their Relations*. Oxford University Press.

Peacocke, C. 1984. Colour concepts and colour experience. *Synthese* 58:365-82.

Peacocke, C. 1989. Perceptual content. In (J.Almog, J. Perry, & H. Wettstein, eds) *Themes from Kaplan*. Oxford University Press.

Peacocke, C. 1992. Scenarios, concepts, and perception. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Peacocke, C. 1994. Nonconceptual content: Kinds, rationales, and relations. *Mind and Language* 4:419-29.

Peacocke, C. 1997. Nonconceptual content defended. *Philosophy and Phenomenological Research*.

Peacocke, C. 2001. Does perception have a nonconceptual content? *Journal of Philosophy* 98:239-264.

Peacocke, C. 2001. Phenomenology and nonconceptual content. *Philosophy and*

Phenomenological Research 62:609-615.

Pendlebury, M. 1987. Perceptual representation. Proceedings of the Aristotelian Society 87:91-106.

Pendlebury, M. 1990. Sense experiences and their contents: A defense of the propositional account. Inquiry 33:215-30.

Lots of reasons why experiences have propositional content (i.e., their content is truth-evaluable, etc). A nice paper.

Schantz, R. 2001. The given regained: Reflections on the sensuous content of experience. Philosophy and Phenomenological Research 62:167-180.

Sedivy, S. 1996. Must conceptually informed perceptual experience involve nonconceptual content? Canadian Journal of Philosophy 26:413-31.

Snowdon, P. 1990. The objects of perceptual experience. Aristotelian Society Supplement, 64:121-50.

Sturgeon, S. 1998. Visual experience. Proceedings of the Aristotelian Society 72:179-200.

Tolhurst, W. 1998. Seemings. American Philosophical Quarterly 35:293-302.

Valberg, J.J. 1992. The Puzzle of Experience. Oxford University Press.

Yoon, B. 2000. Intentionality of perceptual experience. Erkenntnis 52:339-355.

1.5c Representationalism

Beckermann, A. 1995. Visual information-processing and phenomenal consciousness. In (T. Metzinger, ed) Conscious Experience. Ferdinand Schoningh.

Block, N. 1990. Inverted earth. Philosophical Perspectives 4:53-79.
Uses Inverted Earth case, colors and lenses inverted, to argue vs Harman that qualitative states aren't intentional states. Also, less convincingly, to argue that qualitative states aren't functional states.

Block, N. 1996. Mental paint and mental latex. In (E. Villanueva, ed) Perception. Ridgeview.

Block, N. 1999. Sexism, racism, ageism, and the nature of consciousness. Philosophical Topics 26.

Block, N. 2000. Mental paint. In (M. Hahn & M. Ramberg), Essays on Burge. MIT Press.

Dretske, F. 1995. Naturalizing the Mind. MIT Press.

Harman, G. 1990. The intrinsic quality of experience. Philosophical Perspectives.

There are no real qualia problems, just Intentional confusions.

Harman, G. 1996. Explaining objective color in terms of subjective reactions. In (E. Villaneuva, ed) Perception. Ridgeview.

Lalor, B. 1999. Intentionality and qualia. Synthese 121:249-290.

Levine, J. 1997. Are qualia just representations? (Critical notice of Tye) Mind and Language 12:101-13.

Lloyd, D. 1991. Leaping to conclusions: connectionism, consciousness, and the computational mind. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Suggests that conscious states are identical to representational states, and that unconscious representation is impossible; transition between conscious states is non-representational. Appeals to connectionist models in support.

Lloyd, D. 1997. Consciousness and its discontents. *Communication and Cognition* 30:273-284.

Argues that consciousness and representation are distinct, as e.g. the latter depends on context but the former does not.

Lycan, W.G. 1996. Layered perceptual representation. In (E. Villaneuva, ed) *Perception*. Ridgeview.

Lycan, W.G. 1996. *Consciousness and Experience*. MIT Press.

Lycan, W.G. 1998. In defense of the representational theory of qualia. *Philosophical Perspectives* 12:479-87.

McCulloch, G. 1993. The very idea of the phenomenological. *Proceedings of the Aristotelian Society* 67:39-57.

The phenomenological can be reduced to the intentional. Intentional states have a what-it-is-like, and there is no special phenomenal object of introspection.

Neander, K. 1998. The division of phenomenal labor: A problem for representationalist theories of consciousness. *Philosophical Perspectives* 12:411-34.

O'Brien, G. & Opie, J. 1997. Cognitive science and phenomenal consciousness: A dilemma, and how to avoid it. *Philosophical Psychology* 10:269-86.

Rey, G. 1998. A narrow representationalist account of qualitative experience. *Philosophical Perspectives* 12:435-58.

Robinson, W.S. 1998. Intrinsic qualities of experience: Surviving Harman's critique. *Erkenntnis* 47:285-309.

Seager, W.E. 1999. Representational theories of consciousness, parts I and II. In *Theories of Consciousness*. Routledge.

Shoemaker, S. 1990. Qualities and qualia: What's in the mind? *Philosophy and Phenomenological Research Supplement* 50:109-131.

Qualia can't be reduced to standard intentional properties (due to certain inversion cases). Projectivist and sense-reference accounts don't work either. Perhaps qualia are necessarily-illusory intentional properties.

Shoemaker, S. 1991. Qualia and consciousness. *Mind* 100:507-24.

On the relationship between phenomenal and intentional aspects of qualia, and in particular on the accessibility of qualia to conscious awareness. Phenomenal & intentional similarity are connected but must be distinguished.

Stalnaker, R. 1996. On a defense of the hegemony of representation. In (E. Villanueva, ed) *Perception*. Ridgeview.

Sullivan, P.R. 1995. Contentless consciousness and information-processing theories of mind. *Philosophy, Psychiatry, and Psychology* 2:51-59.

Tye, M. 1992. Visual qualia and visual content. In (T. Crane, ed) *The Contents of Experience*. Cambridge University Press.

Tye, M. 1994. Do pains have representational content? In (R. Casati, B.

Smith, & S. White, eds) Philosophy and the Cognitive Sciences.
Holder-Pichler-Tempsky.

Tye, M. 1995. A representational theory of pains and their phenomenal character. *Philosophical Perspectives* 9:223-39.
Argues that pain is representational, and that its phenomenal character is narrow nonconceptual content. They have a complex representational structure, with map-like arrays of sentential contents.

Tye, M. 1995. What "what it is like" is like. *Analysis*.
Argues that "what it is like to be X" is an intentional context, which solves some of the associated problems.

Tye, M. 1996. Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind. MIT Press.

Tye, M. 1996. Orgasms again. In (E. Villanueva, ed) Perception. Ridgeview.

Tye, M. 1997. The problem of simple minds: Is there anything it's like to be a honeybee? *Philosophical Studies* 88:289-317.

Tye, M. 1998. Inverted earth, swampman, and representationalism. *Philosophical Perspectives* 12:459-78.

Stoljar, D. 1996. What what it's like isn't like. *Analysis* 56:281-83.

Vinueza, A. 2000. Sensations and the language of thought. *Philosophical Psychology* 13:373-392.

Wager, A. 1999. The extra qualia problem: Synaesthesia and representationism. *Philosophical Psychology* 12:263-281.

Warfield, T. 1999. Against representational theories of consciousness. *Journal of Consciousness Studies* 6:66-69.

White, S. 1994. Color and notional content. *Philosophical Topics* 22:471-503.

1.5d Internalism and Externalism about Experience [see also 2.2]

Davies, M. 1992. Perceptual content and local supervenience. *Proceedings of the Aristotelian Society* 66:21-45.

Argues that perceptual content does not supervene on internal state, even though it is non-conceptual. Constructs an Twin scenario to that effect.
With remarks on the relation between perceptual content and phenomenology.

Davies, M. 1993. Aims and claims of externalist arguments. In (E. Villanueva, ed) Naturalism and Normativity. Ridgeview.

Distinguishes modal and constitutive externalism, characterizes perceptual content and its relation to sensational content, and argues for externalism about perceptual content by examples.

de Vries, W.A. 1996. Experience and the swamp creature. *Philosophical Studies* 82:55-80.

Argues that a swampthing isn't intelligent or intentional, with different physiological processes and no sensations, as these are functional kinds.

Dretske, F. 1996. Phenomenal externalism, or if meanings ain't in the head, where are qualia? In (E. Villanueva, ed) Perception. Ridgeview.

We only have access to qualia through our concepts, which are external; so internal qualia would be inaccessible. So if qualia are knowable, they're external; and if not, why posit them? With comments by Kim, Horwich, Biro.

Forbes, G. 1997. Externalism and scientific Cartesianism. *Mind and Language* 12:196-205.

Kirk, R. 1994. The trouble with ultra-externalism. *Proceedings of the Aristotelian Society* 68:293-307.

Kirk, R. 1996. Why ultra-externalism goes too far. *Analysis* 56:73-79.

Kirk, R. 1998. Consciousness, information, and external relations. *Communication and Cognition* 30:249-71.

McCulloch, G. 1990. Externalism and experience. *Analysis* 50:244-50.
Argues against McGinn that one should embrace a form of "strong externalism" about experience. Experience can be laden with externally-grounded concepts.

McCulloch, G. 1994. Not much trouble for ultra-externalism. *Analysis* 54:265-9.

Sartwell, C. 1995. Radical externalism concerning experience. *Philosophical Studies* 78:55-70.

There is no epistemically available aspect of experience that is determined internally; experiences are "fused" with the environment.

Tappenden, P. 1996. The roundsquare copula: A semantic internalist's rejoinder. *Proceedings of the Aristotelian Society* 96:395-400.

1.5e Miscellaneous

Cam, P. 1984. Consciousness and content-formation. *Inquiry* 27:381-98.

Carruthers, P. 1998. Conscious thinking: Language or elimination? *Mind and Language* 13:457-476.

Falk, B. 1993. Consciousness, cognition, and the phenomenal. *Proceedings of the Aristotelian Society* 67:55-73.

On conceptual influences on experience, and aspectual seeing, focusing on bodily and dynamic elements. Self-awareness is not of phenomenal states but *in* them. With commentary by S. Mulhall.

Jacquette, D. 1984. Sensation and intentionality. *Philosophical Studies* 47:229-40.

Sensations don't have intentional objects, they *are* intentional objects.

Maloney, J.C. 1986. Sensuous content. *Philosophical Papers* 15:131-54.

McGinn, C. 1988. Consciousness and content. *Proceedings of the British Academy* 74:219-39. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Comparing the problems of consciousness and content, and reconciling optimism on content with pessimism on consciousness. The phenomenological nature of content may be mysterious, but the individuation of contents is not.

Nelkin, N. 1994. Phenomena and representation. *Philosophy of Science* 45:527-47.

Arguing against the view that phenomenal properties are "read off" in making perceptual judgments. Experiences do not literally have color or shape.

Sosa, E. 1986. Experience and intentionality. *Philosophical Topics* 14:67-83.

On a propositional conception of experience, and making sense of awareness of experience and various problems for sense-data monadicism.

1.6 Aspects of Consciousness

1.6a Self-Consciousness [see also 6.20]

- Anscombe, G.E.M. 1975. The first person. In (S. Guttenplan, ed) *Mind and Language*. Oxford University Press.
- Balaban, O. 1990. *Subject and Consciousness: A Philosophical Inquiry into Self-Consciousness*. Rowman & Littlefield.
- Bealer, G. 1997. Self-consciousness. *Philosophical Review* 106:69-117.
- Bermudez, J. 1997. Reduction and the self. *Journal of Consciousness Studies* 4:458-66.
- Bermudez, J. 1998. *The Paradox of Self-Consciousness*. MIT Press.
- Canfield, J.V. 1990. *The Looking-Glass Self: An Examination of Self-Awareness*. Praeger.
- Campbell, J. 1994. *Past, Space, and Self*. MIT Press.
- Campbell, J. 1995. The body image and self-consciousness. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Cassam, Q. 1995. Transcendental self-consciousness, In (P. Kumar, ed) *The Philosophy of P.F. Strawson*. Indian Council for Philosophical Research.
- Cassam, Q. 1997. *Self and World*. Oxford University Press.
- Castaneda, H. 1989. The reflexivity of self-consciousness: Sameness/identity, data for artificial intelligence. *Philosophical Topics* 17:27-58.
- Cheeks, J.M. & Briggs, S.R. 1982. Self-consciousness and aspects of personality. *Journal of Research in Personality* 16:401-8.
- Chisholm, R.M. 1969. On the observability of the self. *Philosophy and Phenomenological Research* 30:7-21.
- Christofidou, A. 2000. Self-consciousness and the double immunity. *Philosophy* 75:539-570.
- Church, J. 1990. Judgment, self-consciousness, and object-independence. *American Philosophical Quarterly* 27:51-60.
- Davis, L.H. 1989. Self-consciousness in chimps and pigeons. *Philosophical Psychology* 2:249-59.
- Delius, H. 1981. *Self-Awareness: A Semantical Inquiry*. Beck.
- Dennett, D.C. 1992. The self as the center of narrative gravity. In (F. Kessel, P. Cole, & D.L. Johnson, eds) *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.
- Eilan, N. 1995. Consciousness and the self. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Eilan, N. Marcel, A.J. & Bermudez, J. 1995. Self-consciousness and the body: Interdisciplinary issues. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Falk, A. 1995. Consciousness and self-reference. *Erkenntnis* 43:151-80.

Frith, U. & Happe, F. 1999. Theory of mind and self-consciousness: What is it like to be autistic? *Mind and Language* 14:1-22.

Gallagher, S. & Meltzoff, A. 1996. The earliest sense of self and others: Merleau-Ponty and recent developmental studies. *Philosophical Psychology* 9:211-33.

Gallagher, S. 1996. The moral significance of primitive self-consciousness: A response to Bermudez. *Ethics* 107:129-40.

Gallagher, S. 2000. Self-reference and schizophrenia: A cognitive model of immunity to error through misidentification. In (D. Zahavi, ed) *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience*. John Benjamins.

Ganerri, J. 2000. Cross-modality and the self. *Philosophy and Phenomenological Research* 61:639-657.

Gennaro, R.J. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333-47.
Argues that consciousness entails having episodic memory, which entails self-consciousness. So consciousness entails self-consciousness.

Gennaro, R.J. 1996. *Consciousness and Self-consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*. John Benjamins.

Hurley, S.L. 1998. Nonconceptual self-consciousness and agency: Perspective and access. *Communication and Cognition* 30:207-247.

James, W. 1890. The consciousness of self. In *The Principles of Psychology*.

Laycock, S.W. 1998. Consciousness it/self. *Journal of Consciousness Studies* 5:141-152.

Martin, M.G.F. 1995. Bodily awareness: A sense of ownership. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.

McCullagh, M. 2000. Functionalism and self-consciousness. *Mind and Language* 15:481-499.

Meijsing, M. 2000. Self-consciousness and the body. *Journal Of Consciousness Studies* 7:34-50.

Metzinger, T. 2000. The subjectivity of subjective experience: A representationist analysis of the first-person perspective. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Mittal, K.K. 1979. Self-identity and self-consciousness. *Indian Philosophical Quarterly* 7:159-63.

Neisser, U. 1991. Five kinds of self-knowledge. *Philosophical Psychology* 1:35-59.

O'Hear, A. 1989. Evolution, knowledge, and self-consciousness. *Inquiry* 32:127-150.

Parker, S.T., Mitchell, R.M., & Boccia, M.L. 1994. *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Proust, J. 2000. Awareness of agency: Three levels of analysis. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Richards, W. 1984. Self-consciousness and agency. *Synthese* 61:149-71.
Self-consciousness is consciousness of agency. Castaneda/Nozick/Nagel.

Shoemaker, S. 1968. Self-reference and self-awareness. *Journal of Philosophy* 65:555-67.

Shoemaker, S. 1986. Introspection and the self. *Midwest Studies in Philosophy*.

Shoemaker, S. 1994. Self-knowledge and "inner sense". *Philosophy and Phenomenological Research* 54:249-314.

Smith, D.W. 1986. The structure of (self-) consciousness. *Topoi* 5:149-56.

Sosa, E. 1983. Consciousness of self and of the present. In (J. Tomberlin, ed) *_Agent, Language, and the Structure of the World_*. Hackett.

Stephens, G. L. & Graham, G. 1994. Self-consciousness, mental agency, and the clinical psychopathology of thought-insertion. *Philosophy, Psychiatry, and Psychology* 1:1-10.

Stephens, G.L. & Graham, G. 2000. *_When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts_*. MIT Press.

Strawson, P.F. 1974. Self, mind, and body. In *_Freedom and Resentment and Other Essays_*.

Strawson, G. 1997. 'The self'. *Journal of Consciousness Studies* 4:405-28.

van Gulick, R. 1988. A functionalist plea for self-consciousness. *Philosophical Review* 97:149-88.

How functionalism can handle consciousness: Self-consciousness is the possession of reflexive metapsychological information. This helps understand learning, representation, and belief. Phenomenal experience is still tricky.

Varela, F.G. 1971. Self-consciousness: Adaptation or epiphenomenon? *Stud Gen* 24:426-39.

White, S. 1987. What is it like to be a homunculus? *Pac Philosophical Quarterly* 68:148-74.

Weird examples of homunculi that are conscious but not self-conscious.
Self-consciousness, not consciousness, is what really counts.

Zahavi, D. 2000. Self and consciousness. In (D. Zahavi, ed) *_Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience_*. John Benjamins.

1.6b The Unity of Consciousness [see also 5.12d, 6.1e, 6.1g]

Arvidson, P. 2000. Transformations in consciousness: Continuity, the self and marginal consciousness. *Journal Of Consciousness Studies* 7:3-26.

Bayne, T. 2000. The unity of consciousness: Clarification and defence. *Australasian Journal Of Philosophy* 78:248-254.

Beahrs, J.O. 1982. *_Unity and Multiplicity: Multilevel Consciousness of Self in Hypnosis, Psychiatric Disorder, and Mental Health_*. Brunner/Mazel.

Beahrs, J.O. 1983. Co-consciousness: A common denominator in hypnosis, multiple personality, and normalcy. *American Journal of Clinical Hypnosis* 26:100-13.

Brooks, D.H.M. 1985. Strawson, Hume, and the unity of consciousness. *Mind* 94:583-86.

Brooks, D.H.M. 1995. *_The Unity of the Mind_*. St. Martin's Press.

- Cotterill, R.M.J. 1995. On the unity of conscious experience. *Journal of Consciousness Studies* 2:290-311.
- Dainton, B. 2000. *_Stream of Consciousness: Unity and continuity in conscious experience_*. Routledge.
- Eccles, J. 1985. *_The Brain and the Unity of Conscious Experience_*. Cambridge University Press.
- Feinberg, T. 2000. The nested hierarchy of consciousness: A neurobiological solution to the problem of mental unity. *Neurocase* 6:75-81.
- Fox, I. 1985. The individualization of consciousness. *Philosophical Topics* 13:119-43.
- Hamlyn, D.W. 1996. The unity of the senses and self-consciousness. In *_Understanding Perception: The Concept and its Conditions_*. Avebury Press.
- Hill, C.S. 1991. Unity of consciousness, other minds, and phenomenal space. In *_Sensations: A Defense of Type Materialism_*. Cambridge University Press.
- Humphrey, N. 2000. One-self: A meditation on the unity of consciousness. *Social Research* 67:1059-1066.
- Hurley, S. 1994. Unity and objectivity. In (C. Peacocke, ed) *_Objectivity, Simulation, and the Unity of Consciousness_*. Oxford University Press.
- Hurley, S. 1998. Unity, neuropsychology, and action. In *_Consciousness in Action_*. Harvard University Press.
- James, W. 1895. The knowing of things together. *Psychological Review* 2:105-24.
- Kobes, B. 2000. Unity of consciousness and bi-level externalism. *Mind and Language* 15:528-544.
- Lockwood, M. 1994. Issues of unity and objectivity. In (C. Peacocke, ed) *_Objectivity, Simulation, and the Unity of Consciousness_*. Oxford University Press.
- Malpas, J. 1999. Constituting the mind: Kant, Davidson, and the unity of consciousness. *International Journal of Philosophical Studies* 7:1-30.
- Marcel, A.J. 1993. Slippage in the unity of consciousness. In *_Experimental and Theoretical Studies of Consciousness_* (Ciba Foundation Symposium 174). Wiley.
- Marcel, A.J. 1994. What is relevant to the unity of consciousness? In (C. Peacocke, ed) *_Objectivity, Simulation, and the Unity of Consciousness_*. Oxford University Press.
- Marks, C. 1980. *_Commissurotomy, Consciousness, and Unity of Mind_*. MIT Press.
- Marks, L.E. 1978. *_The Unity of the Senses: Interrelations among the Modalities_*. Academic Press.
- Maxwell, G. 1978. Unity of consciousness and mind-brain identity. In (J. Eccles, ed) *_Mind and Brain_*. Paragon House.
- McInerney, P.K. 1985. Person-stages and unity of consciousness. *American Philosophical Quarterly* 22:197-209.

Metzinger, T. 1995. Faster than thought: Holism, homogeneity, and temporal coding. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22:396-413. Reprinted in *_Mortal Questions_* (Cambridge University Press, 1979).

Natsoulas, T. 1979. The unity of consciousness. *Behaviorism* 7:45-63.

Natsoulas, T. 1984. Concerning the unity of consciousness: I. Varieties of conscious unity. *Imagination, Cognition and Personality* 3:281-303.

Natsoulas, T. 1986. Concerning the unity of consciousness: II. William James on personal conscious unity. *Imagination, Cognition and Personality* 5:21-30.

Oakley, D.A. & Eames, L.C. 1986. The plurality of consciousness. In (D. Oakley, ed) *_Mind and Brain_*. Methuen.

O'Brien, G. & Opie, J. 1998. The disunity of consciousness. *Australasian Journal of Philosophy* 76:378-95.

O'Brien, G. & Opie, J. 2000. Disunity defended: A reply to Bayne. *Australasian Journal of Philosophy* 78:255-263.

O'Shaughnessy, B. 1994. The diversity and unity of action and perception. In (T. Crane, ed) *_The Contents of Experience_*. Cambridge University Press.

Rosenberg, G.H. 1998. The boundary problem for phenomenal individuals. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness 1996_*. MIT Press.

Rossman, N.I. 1991. *_Consciousness: Separation and Integration_*. SUNY Press.

Schleichert, H. 1985. On the concept of unity of consciousness. *Synthese* 64:411-20.

Shoemaker, S. 1996. Unity of consciousness and consciousness of unity. In *_The First-Person Perspective and Other Essays_*. Cambridge University Press.

Stevenson, L. 2000. Synthetic unities of experience. *Philosophy and Phenomenological Research* 60:281-306.

Tinnin, L. 1990. Mental unity, altered states of consciousness, and dissociation. *Dissociation: Progress in the Dissociative Disorders* 3:154-59.

von der Malsburg, C. 1997. The coherence definition of consciousness. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *_Cognition, Computation, and Consciousness_*. Oxford University Press.

Ward, A. 1980. Materialism and the unity of consciousness. *Analysis* 40:144-46.

Watkins, J.W.N. 1982. A basic difficulty in the mind-brain identity hypothesis. In (J. Eccles, ed) *_Mind and Brain_*. Paragon House.

Zohar, D. 1995. A quantum-mechanical model of consciousness and the emergence of 'I'. *Minds and Machines* 5:597-607.

1.6c Homogeneity of Consciousness (Sellars, etc)

Clark, A. 1989. The particulate instantiation of homogeneous pink. *Synthese* 80:277-304.

Explains homogeneity in terms of nontransitive matching among pixelized parts

of vision. Experience of continuity, not continuous experience. Experiences may have subphenomenal parts (e.g. invisible pixels).

Cornman, J.W. 1970. Sellars, scientific realism, and *sensa*. Review of *Metaphysics* 23:417-51.

Delaney, C.F. 1971. Sellars' grain argument. *Australasian Journal of Philosophy* 50:14-16.

Friedman, I.S. 1989. Ultimate homogeneity: A dialogue. *Philosophy Research Archives* 14:425-53.

Gunderson, K. 1974. The texture of mentality. In (R. Bambrough, ed) *Wisdom: Twelve Essays*. Blackwell.

Lockwood, M. 1993. The grain problem. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

Lycan, W.G. 1987. Sellars' "grain" argument. In *Consciousness*.

Metzinger, T. 1995. Faster than thought: Holism, homogeneity, and temporal coding. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Richardson, R.C. & Muilenberg, G. 1982. Sellars and sense impressions. *Erkenntnis*.

Sellars, W.S. 1963. Philosophy and the scientific image of man. In *Science, Perception, and Reality*. Humanities Press/Ridgeview.

Sellars, W.S. 1971. Seeing, sense impressions, and *sensa*: A reply to Cornman. Review of *Metaphysics*.

1.6d Knowledge of Consciousness [see also 1.3a, 1.8a, 5.13, 6.2i]

Alston, W.P. 1971. Varieties of priveleged access. *American Philosophical Quarterly* 8:223-41.

Alston, W.P. 1983. What's wrong with immediate knowledge? *Synthese* 55:73-96.

Armstrong, D.M. 1963. Is introspective knowledge incorrigible? *Philosophical Review* 62:417-32.

Armstrong, D.M. 1976. Incorrigibility, materialism, and causation. *Philosophical Studies* 30:125-28.

Bayne, T. 2001. Chalmers on the justification of phenomenal judgments. *Philosophy and Phenomenological Research* 62:407-19.

Bradley, R.D. 1964. Avowals of immediate experience. *Mind* 73:186-203.

Chalmers, D.J. 2002. The content and epistemology of phenomenal belief. In (Q. Smith & A. Jokic, eds) *Aspects of Consciousness*. Oxford University Press.

Chandler, J.H. 1970. Incorrigibity and classification. *Australasian Journal of Philosophy* 48:101-6.

Conee, E. 1994. Phenomenal knowledge. *Australasian Journal of Philosophy*.
Mary knew all the facts about qualia beforehand, she just wasn't acquainted with them. Knowledge by acquaintance isn't factual knowledge.

Dretske, F. 1999. The mind's awareness of itself. *Philosophical Studies* 95:103-24.

- Dunlop, C.E.M. 1977. Lehrer and Ellis on incorrigibility. *Australasian Journal of Philosophy* 55:201-5.
- Echelbarger, C.G. 1981. An alleged legend. *Philosophical Studies* 39:227-46.
- Ellis, B. 1976. Avowals are more corrigible than you think. *Australasian Journal of Philosophy* 55:201-5.
- Francescotti, R.M. 1994. Qualitative beliefs, wide content, and wide behavior. *Nous* 28:396-404.
Qualitative beliefs can supervene on behavioral dispositions even if absent/
/inverted qualia are possible. We just individuate belief contents and
behavior widely, with wide content fixed to the qualia.
- Hill, C.S. 1988. Introspective awareness of sensations. *Topoi* 7:11-24.
- Imlay, R.A. 1969. Immediate awareness. *Dialogue* 8:228-42.
- Jackson, F. 1973. Is there a good argument against the incorrigibility thesis? *Australasian Journal of Philosophy* 51:51-62.
- Kirk, R. 1971. Armstrong's analogue of introspection. *Philosophical Quarterly* 21:158-62.
- Kornblith, H. 1998. What is it like to be me? *Australasian Journal of Philosophy* 76:48-60.
- Leeds, S. 1993. Qualia, awareness, Sellars. *Nous* 27:303-330.
A discussion of in what sense we are aware of qualia, and how we can have
beliefs about them, with reference to Sellars. Ends up reducing qualia to
phenomenal beliefs in a language of thought. A rich and subtle paper.
- Levin, J. 2001. The myth of Jones and the return of subjectivity. *Mind and Language* 16:173-192.
- MacDonald, C. 1999. Shoemaker on self-knowledge and inner sense. *Philosophy and Phenomenological Research* 59:711-38.
- Mackie, J.L. 1963. Are there any incorrigible empirical statements? *Australasian Journal of Philosophy* 41:12-28.
- Margolis, J. 1970. Indubitability, self-intimating states, and privileged access. *Journal of Philosophy* 67:918-31.
- Martin, M. 1998. An eye directed outward. In (C. Wright, B. Smith, and C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.
- Nakhtnikian, G. 1968. Incorrigibility. *Philosophical Quarterly* 18:207-15.
- Nida-Rumelin, M. 1995. What Mary couldn't know: Belief about phenomenal states. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.
Reconstructing the Mary case as a Marianna case, and introducing a
distinction between phenomenal and nonphenomenal beliefs, to support the
knowledge argument. A very nice and sophisticated paper.
- Nida-Rumelin, M. 1997. On belief about experiences: An epistemological distinction applied to the knowledge argument. *Philosophy and Phenomenological Research*.
- Odegard, D. 1992. Inner states. *Personalist Forum* 8:265-73.
- Pappas, G. 1974. Incorrigibility, knowledge, and justification. *Philosophical*

Studies 25:219-25.

Pappas, G. 1976. Incorrigibility and central-state materialism. *Philosophical Studies* 29:445-56.

Parsons, K.P. 1970. Mistaking sensations. *Philosophical Review*.

Peacocke, C. 1998. Conscious attitudes, attention, and self-knowledge. In (C. Wright, B. Smith, and C. Macdonald, eds.) *Knowing Our Own Minds*. Oxford University Press.

Pollock, J.L. 1970. Perceptual knowledge. *Philosophical Review* 80:287-319.

Robinson, W.S. 1975. The legend of the given. In (H. Castaneda, ed) *Action, Knowledge, and Reality*. Bobbs-Merrill.

Robinson, W.S. 1982. Causation, sensation, and knowledge. *Mind* 91:524-40.
Argues that we can have non-inferential knowledge of sensations even if they don't make a causal difference to our beliefs, so epiphenomenalism is OK.
All theories have a similar problems.

Rosenberg, J. 2000. Perception vs. inner sense: A problem about direct awareness. *Philosophical Studies* 101:143-160.

Schick, T.W. 1992. The epistemic role of qualitative content. *Philosophy and Phenomenological Research* 52:383-93.
Contra Sellars, Rorty, and Churchland: knowledge of qualitative content is an important aspect of our understanding of mental concepts, although it is not everything.

Sellars, W. 1956. Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science* 1:253-329. Reprinted as *Empiricism and the Philosophy of Mind*. Harvard University Press, 1997.

Shoemaker, S. 1990. First-person access. *Philosophical Perspectives* 4:187-214.
We have a limited special authority about the contents of our mental states. This follows from the link between a state and beliefs about it in the functional definition of that kind of state.

Smart, J.J.C. 1962. Brain processes and incorrigibility. *Australasian Journal of Philosophy* 40:68-70.
Reply to Baier 1962: epistemological authority is compatible with materialism. Mental state reports are not completely incorrigible, though.

Solomon, R.C. 1975. Minimal incorrigibility. *Australasian Journal of Philosophy* 53:254-56.

Sprigge, T.L.S. 1981. Knowledge of subjectivity. *Theoria to Theory* 14:313-25.

Tibbetts, P. 1972. Feigl on raw feels, the brain, and knowledge claims: Some problems regarding theoretical concepts. *Dialectica* 26:247-66.

Tomberlin, J.E. 1975. A problem with incorrigibility. *Philosophia* 5:507-12.

von Eckardt, B. 1975. Some consequences of knowing everything (essential) there is to know about one's mental states. *Review of Metaphysics* 29:3-18.

Wallraff, C.F. 1953. On immediacy and the contemporary dogma of sense-certainty. *Journal of Philosophy*.

Warner, R. 1993. Incorrigibility. In (H. Robinson, ed) *Objections to Physicalism*. Oxford University Press.

Warner, R. 1994. In defense of a dualism. In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.

Warner, R. 1996. Facing ourselves: Incorrighibility and the mind-body problem. *Journal of Consciousness Studies* 3:217-30.

1.6e The Function of Consciousness [see also 1.4e, 6.4a]

Baars, B. 1988. The functions of consciousness. In *A Cognitive Theory of Consciousness*. Cambridge University Press.

Argues for nine major functions of consciousness: in defining inputs, adaptation, debugging, recruiting & control, prioritizing, decision-making, analogy-forming, self-monitoring, and self-maintenance.

Banks, W.P. 1996. How much work can a quale do? *Consciousness and Cognition* 5:368-80.

Bechtel, W. & Richardson, R.C. 1983. Consciousness and complexity: evolutionary perspectives on the mind-body problem. *Australasian Journal of Philosophy* 61:378-95.

Contra Popper, evolution doesn't provide an argument against physicalism or epiphenomenalism. Speculation on what the function of consciousness might be, and how it might be realized: e.g. selecting information.

Block, N. 1995. On a confusion about the function of consciousness. *Behaviora and Brain Sciences* 18:227-47.

Separates phenomenal consciousness from access consciousness, and argues that cases like blindsight only suggest a function for access consciousness, not phenomenal consciousness. The latter remains a mystery. With commentaries.

Bringsjord, S. & Noel, R. 1998. Why did evolution engineer consciousness? In (G. Mulhauser, ed) *Evolving Consciousness*.

DeLancey, C. 1996. Emotion and the function of consciousness. *Journal of Consciousness Studies* 3:492-99.

Dretske, F. 1997. What good is consciousness? *Canadian Journal of Philosophy* 27:1-15.

Flanagan, O. & Polger, T. 1995. Zombies and the function of consciousness. *Journal of Consciousness Studies* 2:313-21.

Argues for the possibility of zombies (contra Moody), then notes that any function could be performed by an unconscious zombie, it seems, so there's no function of consciousness in sight.

Flanagan, O. & Polger, G. 1998. Consciousness, adaptation, and epiphenomenalism. In (G. Mulhauser, ed) *Evolving Consciousness*.

Gregory, R.L. 1996. What do qualia do? *Perception* 25:377-79.

Suggests that qualia serve to distinguish hypotheses about present from past.

Kraemer, E.R. 1984. Consciousness and the exclusivity of function. *Mind* 93:271-5.

Contra Mott 1982: Function needn't be exclusive, and brain processes and consciousness may share a function, due to their close relationship.

McGinn, C. 1981. A note on functionalism and function. *Philosophical Topics* 12:169-70.

Function always underdetermines intrinsic nature, so absent/inverted qualia cases aren't incompatible with consciousness having a function.

Mott, P. 1982. On the function of consciousness. *Mind* 91:423-9.

Consciousness doesn't have a function, as any function it might have is a function of brain processes.

Perlis, D. 1997. Consciousness as self-function. *Journal of Consciousness Studies* 4:509-25.

Place, U. T. 2000. The causal potency of qualia: Its nature and its source. *Brain and Mind* 1:183-192.

Popper, K.R. 1978. Natural selection and the emergence of mind. *Dialectica* 32:339-55.

Ramachandran, V.S. & Hirstein, W. 1998. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies* 4:429-57.

Shanon, B. 1998. What is the function of consciousness? *Journal of Consciousness Studies* 5:295-308.

Tye, M. 1996. The function of consciousness. *Nous* 30:287-305.

Argues that the function of consciousness is not obvious, but that once one accepts a representational view of consciousness, it becomes obvious.

van Gulick, R. 1989. What difference does consciousness make? *Philosophical Topics* 17:211-30.

Trying to counter absent qualia arguments by finding a role for consciousness e.g. in metacognition, or as a way to achieve semantic transparency. But consciousness doesn't seem necessary for these, so it's still a mystery.

van Gulick, R. 1994. Deficit studies and the function of phenomenal consciousness. In (G. Graham & G.L. Stephens, eds) *Philosophical Psychopathology*. MIT Press.

Velmans, M. 1992. Is human information-processing conscious? *Behavioral and Brain Sciences* 14:651-69.

Uses experimental evidence to argue that consciousness is functionally inessential: the tasks associated with consciousness can be performed without consciousness. Only focal-attentive processing is required.

1.7 Qualia

1.7a General

Burgess, J.A. 1990. Phenomenal qualities and the nontransitivity of matching. *Australasian Journal of Philosophy*.

Clark, Andy. 2000. A case where access implies qualia? *Analysis* 60:30-37.

Clark, A. 1985. Qualia and the psychophysical explanation of color perception. *Synthese* 65:377-405.

One can give an information-theoretic explanation of color perception, which leaves nothing out. Rebuts various qualia objections, e.g. from the possibility of inversion. Qualia are codes for external properties.

Clark, A. 1992. *Sensory Qualities*. Clarendon.

Argues that psychology is in the business of explaining sensory qualities, and does a perfectly good job using discriminability as a basis. With detailed argument and many interesting examples.

Clark, A. 2000. *A Theory of Sentience*. Oxford University Press

Cunningham, B. 2001. Capturing qualia: Higher-order concepts and connectionism. *Philosophical Psychology* 14:29-41.

Fox, I. 1989. On the nature and cognitive function of phenomenal content -- Part one. *Philosophical Topics* 17:81-103.

Searching for a theory of qualia: rejects epiphenomenalism, separation of the form and quality of experience, and immediate perception of phenomenal objects. Experience consists in represented (inexistent) objects of thought.

Gilbert, P. 1992. Immediate experience. *Proceedings of the Aristotelian Society* 66:233-250.

Against an account of phenomenal content as given by inner discrimination. Argues that the character of experience consists in its reason-giving role.

Gustafson, D. 1998. Pain, qualia, and the explanatory gap. *Philosophical Psychology* 11:371-387.

Hubbard, T.L. 1996. The importance of a consideration of qualia to imagery and cognition. *Consciousness and Cognition* 3:327-58.

Jakab, Z. 2000. Ineffability of qualia: A straightforward naturalistic explanation. *Consciousness and Cognition* 9:329-351.

Kind, A. 2001. Qualia realism. *Philosophical Studies* 104:143-62.

Kitcher, P.S. 1979. Phenomenal qualities. *American Philosophical Quarterly* 16:123-9.

Qualia problems stem from assuming direct awareness of perceptual states. Instead, we should acknowledge only an ability to detect and label these states. Also argues for the possibility of unconscious and illusory pains.

Leeds, S. 1993. Qualia, awareness, Sellars. *Nous* 27:303-330.

A discussion of in what sense we are aware of qualia, and how we can have beliefs about them, with reference to Sellars. Ends up reducing qualia to phenomenal beliefs in a language of thought. A rich and subtle paper.

Leon, M. 1988. Characterising the senses. *Mind and Language* 3:243-70.

Levine, J. 1995. Qualia: Intrinsic, relational, or what? In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.

Lormand, E. 1995. Qualia! (Now showing at a theater near you.) *Philosophical Topics* 22:127-156.

Mandik, P. 1999. Qualia, space, and control. *Philosophical Psychology* 12:47-60.

Nelkin, N. 1987. How sensations get their names. *Philosophical Studies* 51:325-39.

Sensations are an inessential element of experiences. Experiences are typed by their cognitive component, and the naming of sensations is derivative on this. With examples and empirical evidence about pain, color, perception.

Nelkin, N. 1990. Categorizing the senses. *Mind and Language*.

Putnam, H. 1981. Mind and body. In *_Reason, Truth and History_*. Cambridge University Press.

Considers qualia, inverted and absent, and various other stuff. Wishy-washy.

Rey, G. 1993. Sensational sentences. In (M. Davies & G. Humphreys, eds) *_Consciousness: Philosophical and Psychological Essays_*. Blackwell.

Explicating sensory experience in terms of an appropriate computational relation to a sentence in the language of thought. Argues that this handles

many features of qualia (privacy, ineffability, grainlessness, unity, etc).

Robinson, W.S. 1999. Qualia realism and neural activation patterns. *Journal of Consciousness Studies* 10:65-80.

Shepard, R.N. 1993. On the physical basis, linguistic representation, and conscious experience of colors. In (G. Harman, ed) *Conceptions of the Human Mind: Essays in Honor of George A. Miller*. Lawrence Erlbaum.

Shoemaker, S. 1975. Phenomenal similarity. *Critica* 7:3-37. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).
Where does similarity come from? From belief therein? Similarity of experience = experience of similarity. Also relation to projectibility.

Shoemaker, S. 1994. Phenomenal character. *Nous* 28:21-38.
Phenomenal character is bestowed by representation of certain relational properties, defined by relation to experience. With a discussion of possible candidates, and argument against other views such as projectivism.

Sleutels, J. 1998. Phenomenal consciousness: Epiphenomenalism, naturalism and perceptual plasticity. *Communication and Cognition* 31:21-55.

1.7b Qualia and Materialism [see also 1.2, 1.3]

Clark, A. 1985. A physicalist theory of qualia. *Monist* 68:491-506.
A Goodman-like theory of qualia discrimination.

Cornman, J.W. 1971. *Materialism and Sensations*. Yale University Press.

Double, R. 1985. Phenomenal properties. *Philosophy and Phenomenological Research* 45:383-92.
A somewhat vague defense of materialism against objections from phenomenal properties. The only problems are epistemological.

Harding, G. 1991. Color and the mind-body problem. *Review of Metaphysics* 45:289-307.
On the unique nature of color expanses, which are laid bare to perception as they are in themselves. These are incompatible with functionalist accounts of mind, but might still be physical, on a broader conception thereof.

Holborow, L.C. 1973. Materialism and phenomenal qualities. *Aristotelian Society Supplement* 47:107-19.

Horgan, T. 1987. Supervenient qualia. *Philosophical Review* 96:491-520.
Arguing from the causal efficacy of qualia and the closedness of physical causation to the conclusion that qualia conceptually supervene on the physical. A very thorough paper.

Jolley, K.D. & Watkins, M. 1998. What is it like to be a phenomenologist? *Philosophical Quarterly* 48:204-9.
A reply to Raffman 1995. Maybe our experiences are no more fine-grained than our concepts. Even our experiences of unique hues may be coarse.

Lewis, D. 1995. Should a materialist believe in qualia? *Australasian Journal of Philosophy* 73:140-44.
Materialists can believe in qualia, qua occupier of the folk psychological role. But they cannot accept the Identification Thesis, that having qualia allows us to know exactly what they are.

Lycan, W.G. 1988. Phenomenal objects: A backhanded defense. *Philosophical Perspectives* 3:513-26.
Argues that qualia, if viewed as simple properties of phenomenal individuals,

are problematic for materialism. Considers the case for phenomenal individuals, and argues that they are intentional inexistent.

Marras, A. 1993. Materialism, functionalism, and supervenient qualia. Dialogue 32:475-92.

Qualia aren't reducible to physical properties, but they are supervenient (and ontologically dependent) on microfunctional properties. With remarks on the knowledge argument, Kripke, absent qualia, epiphenomenalism, etc.

Mellor, D.H. 1973. Materialism and phenomenal qualities II. Aristotelian Society Supplement 47:107-19.

Raffman, D. 1995. On the persistence of phenomenology. In (T. Metzinger, ed) Conscious Experience. Ferdinand Schoningh.

Argues that our inability to reidentify qualia is a problem for materialism. How are they represented? Empty demonstrative would be vacuous, predicate would be reidentified, so maybe a plain presentation? Very interesting.

Tye, M. 1986. The subjective qualities of experience. Mind 95:1-17.

Absent/inverted qualia aren't really imaginable. The Knowledge Argument fails, as discovering new experiences doesn't imply learning new facts, but only coming to know old facts in a new way.

1.7c Eliminativism about Qualia [see also 1.4d]

Dennett, D.C. 1978. Why you can't make a computer that feels pain. Synthese 38. Reprinted in Brainstorms* (MIT Press, 1978).

The concept of pain is incoherent, as it's asked to do too many things at once. With a discussion of drugs, flowcharts, reportability, etc.

Dennett, D.C. 1981. Wondering where the yellow went. Monist 64:102-8.

A response to Sellars. All there is to seeing occurrent yellow is the judgment that one is seeing occurrent yellow.

Dennett, D.C. 1988. Quining qualia. In (A. Marcel & E. Bisiach, eds) Consciousness in Contemporary Science. Oxford University Press.

Argues against the existence of ineffable, intrinsic, private, directly accessible properties. With lots of meaty-thought experiments, and arguments that there is no fact of the matter about inversion cases.

Dennett, D.C. 1991. Lovely and suspect qualities. In (E. Villanueva, ed) Consciousness. Ridgeview.

Everett, A. 1996. Qualia and vagueness. Synthese 106:205-226.

There are no qualia: qualia would have to be vague (for Sorites reasons), but there can be no vague properties in nature. The usual Sorites defenses don't work here, as there's no appearance/reality distinction for qualia.

Jacoby, H. 1985. Eliminativism, meaning, and qualitative states. Philosophical Studies 47:257-70.

Arguing against eliminativism for qualia. Even if nothing satisfies all the common-sense properties of qualia, reference of qualia terms is still fixed under a Putnam-style theory of meaning. Argues for scientific functionalism.

Levin, M. 1981. Phenomenal properties. Philosophy and Phenomenological Research 42:42-58.

There are no irreducible phenomenal properties. Materialism can handle our direct awareness of inner states by the right sort of causal connection. Gives a materialism account of discrimination and learning mental concepts.

Levine, J. 1994. Out of the closet: A qualophile confronts qualophobia. Philosophical Topics 22:107-126.

On bold vs. modest qualophilia, and against various qualophobic strategies.
With remarks on scientific objectivity, qualia as an explanandum, and on how
our knowledge of qualia is consistent with the conceivability of zombies.

Ross, D. 1993. Quining qualia Quine's way. Dialogue 32:439-59.

Seager, W.E. 1993. The elimination of experience. Philosophy and
Phenomenological Research 53:345-65.

Dennett's 1988 argument against ineffability, etc., doesn't nearly make the
case against qualia, and largely relies on verificationist assumptions.

Wright, E.W. 1989. Querying "Quining Qualia". Acta Analytica 4:9-32.

1.7d The Inverted Spectrum

Block, N. 1990. Inverted earth. Philosophical Perspectives 4:53-79.

Uses Inverted Earth case, colors and lenses inverted, to argue vs Harman
that qualitative states aren't intentional states. Also, less convincingly,
to argue that qualitative states aren't functional states.

Campbell, N. 2000. Physicalism, qualia inversion, and affective states.
Synthese 124:239-256.

Casati, R. 1990. What is wrong in inverting spectra? Teoria 10:183-6.

Churchland, P.M. & Churchland, P.S. 1981. Functionalism, qualia and
intentionality. Philosophical Topics 12:121-32. Reprinted in _A
Neurocomputational Perspective_ (MIT Press, 1989).

Functional role counts more than qualitative content in determining what
e.g. "redness" is.

Clark, A. 1985. Spectrum inversion and the color solid. Southern Journal of
Philosophy 23:431-43.

Argues that there could be inverted spectra even without a symmetrical color
space. Qualia must be distinguished from their place in color space.

Cole, D.J. 1990. Functionalism and inverted spectra. Synthese 82:207-22.

Acquired spectrum inversions do not refute functionalism, if qualia revert
after behavioral adaptation (as they do with inverting lenses).

Dennett, D.C. 1994. Instead of qualia. In (A. Revonsuo & M. Kamppinen, eds)
Consciousness in Philosophy and Cognitive Neuroscience. Lawrence Erlbaum.

Describes some "inverted spectrum" scenario in computer registers, and argues
that in the absence of a "central clearing house", the inversion of qualia
is indeterminate. There's no reason to believe in non-dispositional qualia.

Gert, B. 1965. Imagination and verifiability. Philosophical Studies 16:44-47.

Inverted spectra with constant behavior is a meaningful hypothesis even under
verificationism. Switching nerve endings, tinting contact lenses, etc.

Hardin, C.L. 1987. Qualia and materialism: Closing the explanatory gap.
Philosophy and Phenomenological Research 48:281-98.

On the physiological bases of phenomenal states, particularly color.
Inverted spectrum isn't really coherent, as coolness/warmth would have to
be inverted too. So the contingency of qualia is diminished.

Hardin, C.L. 1988. _Color for Philosophers_. Hackett.

Distinguishes various functionally distinct inverted spectrum cases.

Hardin, C.L. 1991. Reply to Levine. Philosophical Psychology 4:41-50.

Reply to Levine 1991. "Green residue" and "red residue" may be identical.
Physiology might put more constraints on qualia, eventually ruling out all

other possibilities. But there may still be absent/alien qualia problems.

Hardin, C.L. 1997. Reinverting the spectrum. In (A. Byrne & D.R. Hilbert, eds) *_Readings on Color, Volume 1: The Philosophy of Color_*. MIT Press.

Harrison, B. 1967. On describing colors. *Inquiry* 10:38-52.

Harrison, B. 1973. *_Form and Content_*. Blackwell.
The inverted spectrum is impossible, due to asymmetries in color space.

Harvey, J. 1979. Systematic transposition of colours. *Australasian Journal of Philosophy* 57:211-19.

The inverted spectrum can be detected, if a single person experiences both.

Hatfield, G. 1992. Color perception and neural encoding: Does metameric matching entail a loss of information? *Philosophy of Science Association 1992*, 1:492-504.

Johnsen, B.C. 1986. The inverted spectrum. *Australasian Journal of Philosophy* 64:471-6.

Against Shoemaker: physical realizations do not give empirical conditions for qualia inversion. Nice.

Johnsen, B.C. 1993. The intelligibility of spectrum inversion. *Canadian Journal of Philosophy* 23:631-6.

Kirk, R. 1982. Goodbye to transposed qualia. *Proceedings of the Aristotelian Society* 82:33-44.

The possibility of an inverted spectrum w.r.t. dispositions implies the falsity of physicalism. But this rests on an implausible "slide-viewer" model of seeing, and is incoherent otherwise.

Levine, J. 1988. Absent and inverted qualia revisited. *Mind and Language* 3:271-87.

Inverted qualia, with respect to a functional account, are no more plausible than absent qualia (by analysis of thought experiments). Both lead to first-person skepticism about qualia.

Levine, J. 1991. Cool red. *Philosophical Psychology* 4:27-40.

Contra Hardin 1988: there's a "green residue" after coolness is subtracted, so inverted spectrum could still be possible. In any case, the impossibility of IS doesn't affect the explanatory gap for qualia, which is epistemic.

Lycan, W.G. 1973. Inverted spectrum. *Ratio* 15:315-9.

Inverted spectrum holding behavior constant is at least a coherent idea. Hook up brain in different ways, etc.

Lycan, W.G. 1993. Functionalism and recent spectrum inversions. Manuscript. Argues that qualia are intentional properties, and that inverted spectra, though conceivable, are metaphysically impossible, due to considerations about society and normality. Argues against Block's "inverted earth".

Nida-Rumelin, M. 1996. Pseudonormal vision: An actual case of qualia inversion? *Philosophical Studies* 82:145-57.

A fascinating note on the possibility of people with doubled colorblindness genes, thus inverting color processing; such people may actually exist.

Palmer, S. 1999. Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences*.

Rey, G. 1992. Sensational sentences reversed. *Philosophical Studies* 68:289-319.

Argues for a computational/sentential theory under which qualia are fixed by

functional organization. Argues against Block's 1990 inversion: qualia might slowly change back as associations fade. Memory isn't 100% reliable.

Shoemaker, S. 1975. Phenomenal similarity. *Critica* 7:3-37. Reprinted in *_Identity, Cause, and Mind_* (Cambridge University Press, 1984).

Maybe IS is ongoing, with memory changes. What is the logic of "appears"?

Shoemaker, S. 1982. The inverted spectrum. *Journal of Philosophy* 79:357-381. Reprinted in *_Identity, Cause, and Mind_* (Cambridge University Press, 1984).

All about the coherence and otherwise thereof. Uses switch in state for IS IS wrt behavior. Also claims that IS wrt function is possible as qualia are fixed by realizing state, not functional state. Bad assumption.

Shoemaker, S. 1996. Intersubjective/intrasubjective. In *_The First-Person Perspective and Other Essays_*. Cambridge University Press.

Taylor, D. 1966. The incommunicability of content. *Mind* 75:527-41.

Inverted spectra thought-experiments show that experiential content is incommunicable. Accounts for the fact that attempts to describe such cases lead to contradiction (I'm seeing green & not seeing green).

Tye, M. 1993. Qualia, content, and the inverted spectrum. *Nous*.

Argues that qualia are intentional properties, along the lines of "looks F to P". Handles inverted earth and related cases by taking the narrow intentional content. With remarks on the semantics of color terms.

1.7e Absent Qualia (Block, etc) [see also 1.3b]

Block, N. & Fodor, J.A. 1972. What psychological states are not. *Philosophical Review* 81:159-81.

As a criticism of functionalism. raises the possibility that realizations of any given functional account of mental states may lack qualia.

Block, N. 1980. Troubles with functionalism. In (N. Block, ed) *_Readings in the Philosophy of Psychology_*, Vol 1. Harvard University Press.

All kinds of absent qualia cases: homunculi-headed robots, the population of China, and so on. There is a prima facie doubt that such cases lack qualia, so there is a prima facie case against functionalism.

Bogen, J. 1981. Agony in the schools. *Canadian Journal of Philosophy* 11:1-21.

It's OK for bizarre realizations to lack pain, as functionalism requires teleology as well as organization. With remarks on the relation between pain and "introspectible noxiousness".

Carleton, L. 1983. The population of China as one mind. *Philosophy Research Archives* 9:665-74.

Taking the personal stance, we should regard the Chinese nation as having qualia. A lack of qualia would make a functional difference.

Churchland, P.M. & Churchland, P.S. 1981. Functionalism, qualia and intentionality. *Philosophical Topics* 12:121-32. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).

Absent qualia are impossible. Also, qualia aren't essential to mental state, functional role is.

Cuda, T. 1985. Against neural chauvinism. *Philosophical Studies* 48:111-27.

Replace neurons one by one with homunculi: what happens? Beliefs don't change, does consciousness fade? Very nice.

Elugardo, R. 1983. Functionalism, homunculi-heads and absent qualia. *Dialogue* 21:47-56.

If absent qualia are possible, then either qualia are inexplicable or species

chauvinism is true. Homunculi-heads could make similar arguments about us.

Elugardo, R. 1983. Functionalism and the absent qualia argument. *Canadian Journal of Philosophy* 13:161-80.

Hardcastle, V.G. 1996. Functionalism's response to the problem of absent qualia. *Journal of Consciousness Studies* 3:357-73.

Jacoby, H. 1990. Empirical functionalism and conceivability arguments. *Philosophical Psychology* 2:271-82.

Conceivability arguments are only a problem for empirical functionalism insofar as they are a problem for materialism in general. Very true.

Juhl, C.F. 1998. Conscious experience and the nontriviality principle. *Philosophical Studies* 91:91-101.

Levin, J. 1985. Functionalism and the argument from conceivability. *Canadian Journal of Philosophy Supplement* 11:85-104.

Argues that metaphysical conclusions can be drawn from conceivability arguments, but that absent qualia cases have not been clearly and distinctly conceived. The functionalist is better off than the identity theorist here.

Levine, J. 1988. Absent and inverted qualia revisited. *Mind and Language* 3:271-87.

IQ are no more plausible than AQ (by analysis of thought experiments and skepticism). So there's no reason to choose physicalist-functionalism over pure functionalism, as Shoemaker does. Nice.

Sayan, E. 1988. A closer look at the Chinese Nation argument. *Philosophy Research Archives* 13:129-36.

The Chinese Nation would require less people than Churchland & Churchland 1981 suggest, as we'd only need to handle a subset of all possible inputs.

Tye, M. 1993. Blindsight, the absent qualia hypothesis, and the mystery of consciousness. In (C. Hookway, ed) *Philosophy and the Cognitive Sciences*. Cambridge University Press.

Gives a thorough neurophysiological analysis of blindsight and related pathologies, and argues that these cannot be used to support the possibility of absent qualia. With remarks on the mystery of consciousness.

1.7f Introspection and Absent Qualia (Shoemaker) [see also 1.6d]

Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:291-315. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Absent qualia possible => qualia make no causal difference => no knowledge of qualia, therefore absent qualia are impossible. If qualia are introspectively accessible, they must be functional. An important argument.

Shoemaker, S. 1981. Absent qualia are impossible -- A reply to Block. *Philosophical Review* 90:581-99. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Reply to Block 1980. Distinguishes two AQ theses, and argues that if AQ are possible, then the problem for functionalism isn't due solely to qualia.

Averill, E.W. 1990. Functionalism, the absent qualia objection, and eliminativism. *Southern Journal of Philosophy* 28:449-67.

Defending Shoemaker's argument against Conee: immediate awareness and qualitative beliefs are the same. But maybe people **can't** tell whether they're having genuine or ersatz pain. Eliminativism is the best option.

Block, N. 1980. Are absent qualia impossible? *Philosophical Review* 89:257-74.

Reply to Shoemaker 1975. The possibility of absent qualia is compatible with a functional role for qualia, as qualia can make a causal difference that is independent of a given functional account.

Conee, E. 1985. The possibility of absent qualia. *Philosophical Review* 94:345-66.

Contra Shoemaker: qualia cause qualitative beliefs, which are affected by the absence of qualia, so we know about qualia even if AQ are possible.

Davis, L. 1982. Functionalism and absent qualia. *Philosophical Studies* 41:231-49.

Elucidating Shoemaker's argument: if absent qualia are possible, then the difference between real and ersatz pain makes no difference to belief, so qualia aren't introspectively accessible. A nice analysis.

Doore, G. 1981. Functionalism and absent qualia. *Australasian Journal of Philosophy* 59:387-402.

Qualia and qualitative beliefs are the same, so Shoemaker's argument fails. A numbness/pain inversion argument shows that pain isn't a functional state; it yields an introspectible difference without a functional difference.

Francescotti, R.M. 1994. Qualitative beliefs, wide content, and wide behavior. *Nous* 28:396-404.

Qualitative beliefs can supervene on behavioral dispositions even if absent/inverted qualia are possible. We just individuate belief contents and behavior widely, with wide content fixed to the qualia.

Hill, C.S. 1991. Introspection and the skeptic. In *Sensations: A Defense of Type Materialism*. Cambridge University Press.

Argues that the possibility of absent qualia is compatible with introspective knowledge. The fact that we have evidence of qualia isn't altered by the fact that we'd still think we had that evidence if we didn't have qualia.

White, N. 1985. Professor Shoemaker and the so-called 'qualia' of experience. *Philosophical Studies* 47:369-383.

Shoemaker's account leaves out experienced relations, such as experienced similarity. Experienced similarity is not the same as similarity between experiences. Being experienced is not an experienced feature.

1.7g Functionalism and Qualia, General

Brown, M. 1983. Functionalism and sensations. *Auslegung* 10:218-28.

Various comments on functionalism's troubles with qualia, including absent and inverted qualia. Analogis with biology and information theory.

Chalmers, D.J. 1995. Absent qualia, fading qualia, dancing qualia. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh. Press.

Argues that absent qualia and inverted qualia are empirically impossible (though logically possible), using neural-replacement thought-experiments. So functional organization fully determines conscious experience.

Cole, D.J. 1990. Functionalism and inverted spectra. *Synthese* 82:207-22.

Acquired spectrum inversion doesn't refute functionalism, if qualia revert after behavioral adaptation. With empirical evidence.

Dumpleton, S. 1988. Sensation and function. *Australasian Journal of Philosophy* 66:376-89.

Eshelman, L.J. 1977. Functionalism, sensations, and materialism. *Canadian Journal of Philosophy* 7:255-74.

Graham, G. & Stephens, G. 1985. Are qualia a pain in the neck for functionalists? *American Philosophical Quarterly* 22:73-80.

Pain-qualia are in the body, not the mind, and so aren't part of psychology.

Graham, G. & Stephens, G. 1987. Minding your P's and Q's: Pain and sensible qualities. *Nous* 21:395-405.

Greenberg, W. 1998. On Chalmers' "principle of organizational invariance" and his "dancing qualia" and "fading qualia" thought experiments. *Journal of Consciousness Studies* 5:53-58.

Hill, C.S. 1991. The failings of functionalism. In *_Sensations: A Defense of Type Materialism_*. Cambridge University Press.

Gives a number of arguments against both analytic functionalism and psychofunctionalism: arguments from absent qualia, absent functional role, epistemology, semantics, and heterogeneity of functional roles.

Horgan, T. 1984. Functionalism, qualia, and the inverted spectrum. *Philosophy and Phenomenological Research*, 44:453-69.

Argues that non-phenomenal mental events are functional, while qualia are low-level physiological.

Jarrett, G. 1996. Analyzing mental demonstratives. *Philosophical Studies* 84:49-62.

Lycan, W.G. 1981. Form, function and feel. *Journal of Philosophy* 78:24-50.

Accuses Block of a perspective error. Functionalism can handle a lot, if it's multi-levelled.

Lycan, W.G. 1987. Homunctionalism and qualia. In *_Consciousness_*. MIT Press.

Various stuff, mostly against absent qualia arguments.

Moor, J.H. 1988. Testing robots for qualia. In (H. Otto & J. Tuedio, eds) *_Perspectives on Mind_*. Kluwer.

Behavioral evidence for qualia is always indirect. And you can't check by replacing own neurons by chips, as you'll still believe you have qualia if you're functionally identical. Posit robot qualia as explanatory construct?

Nemirow, L. 1979. Functionalism and the subjective quality of experience. Dissertation, Stanford University.

Rey, G. 1994. Wittgenstein, computationalism, and qualia. In (R. Casati, B. Smith, & S. White, eds) *_Philosophy and the Cognitive Sciences_*.

Holder-Pichler-Tempsky.

Computational functionalism about qualia is compatible with Wittgenstein's views. It makes sense of the points about "dividing through" my private objects, for example. With remarks on spectrum inversions.

Seager, W.E. 1983. Functionalism, qualia and causation. *Mind* 92:174-88.

Functionalism can't explain the causal role of qualia by identifying them with functional states (circularity) or physical realizations (chauvinism). Which leaves property dualism, epiphenomenalism, or eliminativism for qualia.

Shoemaker, S. 1994. The first-person perspective. *Proceedings and Addresses of the American Philosophical Association* 68:7-22.

Against drawing strong conclusions from first-person imaginings. Considers Searle's silicon-replacement scenario: we might infer that perception isn't veridical, that there's another mind about, or even another body.

van Gulick, R. 1988. Qualia, functional equivalence and computation. In (H. Otto & J. Tuedio, eds) *_Perspectives on Mind_*. Kluwer.

Commentary on Moor 1988. Systems that differ in qualitative properties will

likely differ in functional organization.

van Heuveln, B., Dietrich, E. & Oshima, M. 1998. Let's dance! The equivocation in Chalmers' dancing qualia argument. Minds and Machines.

White, S. 1986. Curse of the qualia. Synthese 68:333-68.
Criticism of "physicalist-functionalism", where functional organization doesn't completely determine qualia (e.g. Shoemaker/Block). The only tenable options are pure functionalism or transcendental dualism. Nice.

White, S. 1989. Transcendentalism and its discontents. Philosophical Topics 17:231-61.
Taking transcendental dualism seriously. Privileged access provides strong arguments against objective theories, but it turns out that transcendentalism can't explain it any better, so maybe embrace objective theories after all.

Wright, E. 1995. More qualia trouble for functionalism: The Smythies TV-hood analogy. Synthese 97:365-82.

Zuboff, A. 1994. What is a mind? Midwest Studies in Philosophy 19:183-205.
Replacing a brain chunk while preserving causal role must preserve experience;

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 2: Mental Content [869]

=====

2.1 Propositional Attitudes and Folk Psychology [285]

2.1a The Language of Thought (Fodor) [54]

2.1b The Intentional Stance (Dennett) [40]

2.1c Eliminativism (Churchlands, etc) [57]

2.1d Propositional Attitudes, General [42]

2.1e The Nature of Folk Psychology [38]

2.1f The Simulation Theory [54]

2.2 Internalism and Externalism [278]

2.2a Is Content in the Head? (Putnam, Burge) [38]

2.2b Externalism and Psychological Explanation [41]

2.2c Externalism and Mental Causation [28]

2.2d Externalism and the Theory of Vision [16]

2.2e Externalism and Computation [11]

2.2f Externalism and Self-Knowledge [72]

2.2g The Status of Narrow Content [43]

2.2h Miscellaneous [29]

2.3 Theories of Content [144]

2.3a Information-Based Accounts (Dretske, etc) [29]

2.3b Asymmetric Dependence (Fodor) [20]

2.3c Causal Accounts, General [17]

2.3d Teleological Approaches (Millikan, etc) [48]

2.3e Conceptual Role Approaches [19]

2.3f Theories of Content, Misc [11]

2.4 Representation, General [25]

2.5 The Explanatory Role of Content (Dretske, etc) [22]

2.6 Concepts [46]

2.7 Meaning Holism [26]

2.8 Mental Content, Misc [43]

2.1 The Status of Propositional Psychology

2.1a The Language of Thought (Fodor)

Fodor, J.A. 1975. The Language of Thought. Harvard University Press.
Argues that thought involves computation upon representations, and that these are structured as sentences in a mental language. With linguistic and psychological evidence, and arguments that the mental language is innate.

Fodor, J.A. 1987. Why there still has to be a language of thought. In Psychosemantics. MIT Press.
Because it fits explanatory methodology, it coheres with the usual ontology of psychological processes, and it explains systematicity.

Fodor, J.A. 1978. Propositional attitudes. *Monist* 61:501-23. Reprinted in RePresentations (MIT Press, 1980).
About what PA's are, and why they're at the foundations of thought.

Fodor, J. 2001. Language, thought and compositionality. *Mind and Language* 16:1-15.

Abbott, B. 1995. Natural language and thought: Thinking in English. *Behavior and Philosophy* 23:49-55.

Bonjour, L. 1991. Is thought a symbolic process? *Synthese* 89:331-52.
Argues that symbol processing can't account for the intrinsically contentful nature of thought: using a symbol doesn't give understanding of its content. With defense against arguments from twin earth and conceptual-role semantics.

Braddon-Mitchell, D. & Fitzpatrick, J. 1990. Explanation and the language of thought. *Synthese* 83:3-29.
No need to postulate LOT: diachronic explanation is as good as synchronic, and high-level laws can exist without high-level causal connections.

Clapin, H. 1997. Problems with principle P. *Pacific Philosophical Quarterly* 78:261-??.

Clark, A. 1988. Thoughts, sentences and cognitive science. *Philosophical Psychology* 1:263-78.

Crane, T. 1990. The language of thought: No syntax without semantics. *Mind and Language* 5:187-213.

Davies, M. 1992. Aunty's own argument for the language of thought. In (J. Ezquerro & J. Larrazabal, eds) Cognition, Semantics and Philosophy. Kluwer.

Dennett, D.C. 1977. A cure for the common code. *Mind*. Reprinted in Brainstorms (MIT Press, 1978).
Review of Fodor's LOT. Fodor's view is too strong: function, not structure, is criterial for content. The structure of a predictive theory need not be directly reflected in inner processing.

Dennett, D.C. 1975. Brain writing and mind reading. *Minnesota Studies in the Philosophy of Science* 7:403-15. Reprinted in Brainstorms (MIT Press, 1978).
On the explicit representation of belief: criteria, plausibility, and relationship to verbal reports and conscious judgments.

Dennett, D.C. 1990. Granny's campaign for safe science. In (B. Loewer & G. Rey, eds) Meaning in Mind: Fodor and his Critics. Blackwell.
A general treatment of Fodor, identifying him as arch-conservative mentalist.

DeWitt, R. 1995. Vagueness, semantics, and the language of thought. *Psyche* 1.

Dunlop, G. 1990. Conceptual dependency as the language of thought. *Synthese*

82:275-96.

Relates Schank's conceptual dependency to Fodor's LOT.

Egan, M.F. 1991. Propositional attitudes and the language of thought.
Canadian Journal of Philosophy 21:379-88.

Contra two of Fodor's arguments for LOT. Complex causes need not have LOT constituency structure; and evidence from psychological theory falls short.

Field, H. 1978. Mental representation. Erkenntnis 13:9-18. Reprinted in
(N. Block, ed) *_Readings in the Philosophy of Psychology_* (MIT Press, 1980).
Analyzes belief into a relation between a person and an internal sentence,
along with a semantic relation between that sentence and e.g. a proposition.
With arguments against functionalist analyses, and against propositions.

Garson, J.W. 1997. Syntax in a dynamic brain. Synthese 110:343-55.
There are no good arguments for LOT of the form "The brain needs to do X, and
X entails LOT". Considers X = concatenation, logical form, tracking,
combinatorial encoding. Either LOT is weakened deeply or is unnecessary.

Garfield, J. 2000. Thought as language: A metaphor too far. Protosociology
14:85-101.

Gauker, C. 1995. *_Thinking Out Loud: An Essay on the Relation between Thought
and Language_*. Princeton University Press.

Harman, G. 1973. *_Thought_*. Princeton University Press.

Harman, G. 1975. Language, thought, and communication. Minnesota Studies in
the Philosophy of Science 7:270-298.
Argues that the primary role of language is in thought rather than in
communication, and the language of thought incorporates natural language.

Harman, G. 1977. How to use propositions. American Philosophical Quarterly.

Harman, G. 1978. Is there mental representation? Minnesota Studies in the
Philosophy of Science 9.

Hauser, L. 1995. Natural language and thought: Doing without mentalese.
Behavior and Philosophy 23:41-47.

Heil, J. 1981. Does cognitive psychology rest on a mistake? Mind 90:321-42.
LOT confuses processes with descriptions of processes. Also, symbols cannot
denote solely in virtue of structure, so must rely on human interpretation.

Kaye, L.J. 1994. The computational account of belief. Erkenntnis 40:137-53.

Kaye, L.J. 1995. The languages of thought. Philosophy of Science 62:92-110.

Knowles, J. 1998. The language of thought and natural language
understanding. Analysis 58:264-272.

Loar, B. 1982. Must beliefs be sentences? Philosophy of Science Association.

Lycan, W.G. 1982. Toward a homuncular theory of believing. Cognition and
Brain Theory 4:139-59.
Defends sententialism of the homuncular variety: little modules all the
way in. Lots of pro-belief arguments.

Lycan, W.G. 1990. Mental content in linguistic form. Philosophical Studies
58:147-54.
Distinguishes varieties of Sententialism, reasonable vs. mad-dog.

Lycan, W.G. 1993. A deductive argument for the representational theory of
thinking. Mind and Language 8:404-22.

Argues from the unboundedness of thinking and the need for a finite stock of elements to something like a language of thought. With remarks on connectionism and instrumentalism, and a reply by Stalnaker.

Laurence, S. & Margolis, E. 1997. Regress arguments for the language of thought. *Analysis* 57:60-66.

Marras, A. 1987. The weak and the strong representational theory of mind: Stich's interpretation of Fodor. *Dialogue* 26:349-55.

Matthews, R.J. 1989. The alleged evidence for representationalism. In (S. Silvers, ed) *_Rerepresentation_*. Kluwer.

Argues that contrary to some claims, cognitive psychology does not provide much support for a computational/representational theory of propositional attitudes. Specifically considers research in psycholinguistics and vision.

Matthews, R.J. 1991. Is there vindication through representationalism? In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell. Fodor's theory can't deal with inexplicit attitudes: the core/derivative distinction is untenable. But we can make sense of intentional causation and psychological explanation without explicit representation.

Millikan, R.G. 1993. On mentalese orthography. In (B. Dahlbom, ed) *_Dennett and his Critics_*. Blackwell.

On some problems typing tokens in the language of thought. There's no principled distinction between type-identical tokens and type-distinct tokens with an identity judgment. With interesting remarks on co-identification.

Pessin, A. 1995. Mentalese syntax: Between a rock and two hard places. *Philosophical Studies* 78:33-53.

Argues that there is no good way to individuate syntactic types in Mentalese. Neural typing, causal typing, and semantic typing all fail.

Rantala, V. & Vaden, T. 1997. Minds as connoting systems: Logic and the language of thought. *Erkenntnis* 46:315-334.

Rey, G. 1995. A not "merely empirical" argument for the language of thought. *Philosophical Perspectives* 9:201-22.

Schiffer, S. 1991. Does Mentalese have a compositional semantics? In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell.

Argues that the language of thought need not have a compositional semantics; productivity and systematicity can be explained without it.

Schiffer, S. 1994. The language-of-thought relation and its implications. *Philosophical Studies* 76:263-85.

Schwartz, G. 1996. Symbols and thought. *Synthese* 106:399-407.

Sher, G. 1975. Sentences in the brain. *Philosophy and Phenomenological Research* 36:94-99.

On Danto's suggestion that beliefs are like sentences. Conventionality poses problems, as does differentiating between different sorts of attitudes.

Stalnaker, R.C. 1990. Mental content and linguistic form. *Philosophical Studies* 58:129-46.

Sterelny, K. 1983. Mental representation: What language is Brainsese? *Philosophical Studies*, 43:365-82.

Motivates LOT and defends it against various objections: e.g. tacit belief, identity conditions, infinite regress, and semantic nativism.

Stich, S.P. 1978. Beliefs and subdoxastic states. *Philosophy of Science*

45:499-518.

Teng, N.Y. 1999. The language of thought and the embodied nature of language use. *Philosophical Studies* 94:237-251.

Tienson, J. 1990. Is this any way to be a realist? *Philosophical Psychology*.

Warmbrod, K. 1989. Beliefs and sentences in the head. *Synthese* 2:201-30.

Weller, C. 1997. Bonjour and mentalese. *Synthese* 113:251-63.

Yagisawa, T. 1994. Thinking in neurons: Comments on Stephen Schiffer's "The language-of-thought relation and its implications". *Philosophical Studies* 76:287-96.

2.1b The Intentional Stance (Dennett)

Dennett, D.C. 1978. *_Brainstorms_*. MIT Press.

Dennett, D.C. 1971. Intentional systems. *Journal of Philosophy* 68:87-106
Reprinted in *_Brainstorms_* (MIT Press, 1978).

Can view systems from physical stance, design stance, or intentional stance.
Beliefs/desires are attributed under the intentional stance, with help from
certain idealized norms of rationality and accuracy licensed by evolution.

Dennett, D.C. 1981. Making sense of ourselves. *Philosophical Topics* 12:63-81.
Reprinted in *_The Intentional Stance_* (MIT Press, 1987).

Reply to Stich 1981. Irrationality is misdesign (take design stance). Etc.

Dennett, D.C. 1987. *_The Intentional Stance_*. MIT Press.

Beliefs/desires are useful predictive attributions. This isn't inconsistent
with a certain degree of realism (abstracta/illata distinction).

Dennett, D.C. 1988. *Precis of _The Intentional Stance_*. Behavioral and Brain
Sciences.

TIS, with commentaries and replies.

Dennett, D.C. 1990. The interpretation of texts, people and other artifacts.
Philosophy and Phenomenological Research (Supplement) 50.

Mental states are underdetermined: like interpreting a text, or finding an
object's function. Even adaptationist teleology gives no fact of the matter.

Dennett, D.C. 1991. Real patterns. *Journal of Philosophy* 88:27-51.

Proposition attitudes have the ontological status of a noisy pattern that
helps make sense of behavior. This degree of realism falls on a scale:
Fodor > Davidson > Dennett > Rorty > Churchland.

Baker, L.R. 1987. Instrumentalism: Back from the brink? In *_Saving Belief_*.
Princeton University Press.

Dennett vacillates between stance-dependence, -independence; e.g. on
rationality, design features. Instrumentalism can't be rescued.

Baker, L.R. 1989. Instrumental intentionality. *Philosophy of Science*
56:303-16.

Bechtel, W. 1985. Realism, instrumentalism, and the intentional stance.
Cognitive Science 9:265-92.

Dennett should be a realist, of the relative-to-environment variety.

Byrne, A. 1998. Interpretivism. *European Review of Philosophy* 3.

Cam, P. 1984. Dennett on intelligent storage. *Philosophy and Phenomenological
Research* 45:247-62.

- Clark, A. 1990. Belief, opinion and consciousness. *Philosophical Psychology*. Argues contra Dennett and Smolensky that language is fundamental, not just an add-on.
- Cohen, B. 1995. Patterns lost: Indeterminism and Dennett's realism about beliefs. *Pacific Philosophical Quarterly* 76:17-31.
- Cummins, R. 1981. What can be learned from *_Brainstorms_*? *Philosophical Topics* 12:83-92.
Questioning Dennett on the bridge between intentional characterization and functional characterization. Arguing for the importance of context.
- Davies, D. 1995. Dennett's stance on intentional realism. *Southern Journal of Philosophy* 33:299-312.
- Fodor, J.A. 1981. Three cheers for propositional attitudes. In *_Representations_*. MIT Press.
Dennett's rationality/intentional idealization assumptions should not be viewed as Platonic but epistemic. PA's are real and play real roles.
- Fodor, J.A. & LePore, E. 1993. Is intentional ascription intrinsically normative? In (B. Dahlbom, ed) *_Dennett and His Critics_*. Blackwell.
Against "interpretivism" about intentionality: projectivism is hopeless, and Dennett's arguments for normativism (via charity and evolution) go wrong or beg the question.
- Foss, J. 1994. On the evolution of intentionality as seen from the intentional stance. *Inquiry* 37:287-310.
- Gauker, C. 1988. Objective interpretationism. *Pacific Philosophical Quarterly* 69:136-51.
- Haugeland, J. 1993. Pattern and being. In (B. Dahlbom, ed) *_Dennett and His Critics_*. Blackwell.
- Heitner, R. 2000. Is design relative or real? Dennett on intentional relativism and physical realism. *Minds and Machines* 10:267-83.
- Hornsby, J. 1992. Physics, biology, and common-sense psychology. In (D. Charles & K. Lennon, eds) *_Reduction, Explanation and Realism_*. Oxford University Press.
- Kukla, R. 2000. How to get an interpretivist committed. *Protosociology* 14:180-221.
- Lyons, W. 1990. Intentionality and modern philosophical psychology, I. The modern reduction of intentionality. *Philosophical Psychology* 3:247-69.
- McLaughlin, B. & O'Leary-Hawthorne, J. 1995. Dennett's logical behaviorism. *Philosophical Topics* 22:189-258.
- McLaughlin, B. 2000. Why intentional systems theory cannot reconcile physicalism with realism about belief and desire. *Protosociology* 14:145-157.
- McCulloch, G. 1990. Dennett's little grains of salt. *Philosophical Quarterly* 40:1-12.
Dennett must be one of: realist, eliminativist, instrumentalist.
- Narayanan, A. 1996. The intentional stance and the imitation game. In (P. Millican & A. Clark, eds) *_Machines and Thought_*. Oxford University Press.
- Nelkin, N. 1993. Patterns. *Mind and Language* 9:56-87.

Dennett's instrumentalism can't explain the acquisition of intentional concepts. Proposition attitudes are directly introspectible entities, although still theoretical and still patterns.

Price, H. 1995. Psychology in perspective. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind*. Kluwer.

Radner, D. & Radner, M. 1995. Cognition, natural selection, and the intentional stance. *International Studies in the Philosophy of Science* 9:109-19.

Richard, M. 1995. What isn't a belief? *Philosophical Topics* 22:291-318.

Richardson, R.C. 1980. Intentional realism or intentional instrumentalism? *Cognition and Brain Theory* 3:125-35.

Sharpe, R. 1989. Dennett's journey towards panpsychism. *Inquiry* 32:233-40.

Slors, M. 1996. Why Dennett cannot explain what it is to adopt the intentional stance. *Philosophical Quarterly* 46:93-98.

Stich, S.P. 1980. Headaches. *Philosophical Books* 21:65-73.
Critical review of *Brainstorms*, with response.

Stich, S.P. 1981. Dennett on intentional systems. *Philosophical Topics* 12:39-62. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).
Dennett has problems with rationality, realism, etc. Hard line/soft line: either intentional stance is too close to FP or too far away.

Webb, S. 1994. Witnessed behavior and Dennett's intentional stance. *Philosophical Topics* 22:457-70.

Yu, P. & Fuller, G. 1986. A critique of Dennett. *Synthese* 66:453-76.
Very thorough account of the evolution of Dennett's views. Elucidates abstracta/illata, criticizes intentional subpersonal psychology.

2.1c Eliminativism (Churchlands) [see also 4.3c]

Churchland, P.S. 1980. Language, thought, and information processing. *Nous* 14:147-70.

Sentential processing is out. Against Harman's mental English and Fodor's Mentalese. Arguments from learning, evolution, neuroscience, mental images.

Churchland, P.M. 1981. Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78:67-90. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Eliminate beliefs/desires, remnants of a stagnant folk theory.

Churchland, P.M. & Churchland, P.S. 1983. Stalking the wild epistemic engine. *Nous* 17:5-20. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

How to dethrone language and still handle content.

Churchland, P.M. 1985. On the speculative nature of our self-conception. *Canadian Journal of Philosophy Supplement* 11:157-173.

Reply to Foss 1985: EM is plausible, though certainly not applicable everywhere -- e.g. sensations will be reduced, not eliminated.

Churchland, P.M. 1989. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT Press.

14 glimpses of the neurophilosophical golden age.

Churchland, P.M. 1993. Theory, taxonomy, and methodology: A reply to Haldane's "Understanding folk". Proceedings of the Aristotelian Society 67:313-19.

Reply to Haldane 1988. Even observations can be reconceived. With remarks perceptual plasticity and propositions, and a rejoinder by Haldane.

Churchland, P.M. 1993. Evaluating our self-conception. Mind and Language 8:211-22.

It's "bad faith" to accept modern epistemology but to deny the possibility of eliminativism. On various objections: "functional kinds", "self-defeating", "what could falsify it?", "different purposes", "no alternatives".

Baker, L.R. 1987. The threat of cognitive suicide. In Saving Belief. Princeton University Press.

Elaborating the paradoxes of disbelieving in belief. Rational acceptability, assertion, and truth are all at risk.

Baker, L.R. 1988. Cognitive suicide. In (R. Grimm & D. Merrill, eds) Contents of Thought. University of Arizona Press.

Eliminativism is pragmatically incoherent, as it implies that language isn't meaningful and that the thesis isn't formulable. Folk psychology needn't be scientifically reduced to be true. With comments by Chastain, and reply.

Bertolet, R. 1994. Saving eliminativism. Philosophical Psychology 7:87-100. Against Baker's cognitive-suicide arguments against eliminativism. We don't know what a replacement theory will look like, but that doesn't show that none is forthcoming.

Bickle, J. 1992. Revisionary physicalism. Biology and Philosophy 7:411-30. Argues for a revisionary reduction of the propositional attitudes, rather than elimination or smooth reduction. Sentential aspects will go, but coarse-grained functional profiles and content will remain.

Blunt, P.K. 1992. A defense of folk psychology. International Philosophical Quarterly 32:487-98.

Boghossian, P. 1990. The status of content. Philosophical Review 99:157-84. Irrealism about mental content (and therefore truth-conditions) can't be made sense of. An error thesis presupposes factual truth-conditions, and a non-factualist thesis presupposes a non-deflationary theory of truth.

Boghossian, P. 1991. The status of content revisited. Pacific Philosophical Quarterly 71:264-78. Reply to Devitt 1990.

Chater, N. & Oaksford, M. 1996. The falsity of folk theories: Implications for psychology and philosophy. In (W. O'Donahue & R. Kitchener, eds) The Philosophy of Psychology. Sage Publications.

Clark, A. 1996. Dealing in futures: Folk psychology and the role of representations in cognitive science. In (R. McCauley, ed) The Churchlands and their Critics. Blackwell.

Cling, A. 1989. Eliminative materialism and self-referential inconsistency. Philosophical Studies 56:53-75. Unbelief in belief is not incoherent. Argues with Baker.

Cling, A. 1990. Disappearance and knowledge. Philosophy of Science 57:226-47.

Cling, A. 1991. The empirical virtues of belief. Philosophical Psychology 4:303-23. Cognitive states like belief are necessary to explain the dependence of behavior on perceptual features of the environment. Informational states

alone are not enough, as they can't explain selective response to features.

Devitt, M. 1990. Transcendentalism about content. *Pacific Philosophical Quarterly* 71:247-63.

Against Boghossian's critique: the eliminativism will express her claim in a new framework, so appeals to truth beg the question. With a response.

Devitt, M. & Rey, G. 1991. Transcending transcendentalism. *Pacific Philosophical Quarterly* 72:87-100.

Rejoinder to Boghossian 1990.

Foss, J.E. 1985. A materialist's misgivings about eliminative materialism. *Canadian Journal of Philosophy Supplement* 11:105-33.

EM needs much more evidence before being so gung ho.

Graham, G. & Horgan, T. 1992. Southern fundamentalism and the end of philosophy. In (E. Villanueva, ed) *_Truth and Rationality_*. Ridgeview.

Greenwood, J.D. 1991. Reasons to believe. In (J. Greenwood, ed) *_The Future of Folk Psychology_*. Cambridge University Press.

Argues that folk psychological states exist, even if they aren't useful as causal explanation. We have independent reason to believe in them, e.g. from self-knowledge. They're useful in social psychology, too.

Greenwood, J.D. 1992. Against eliminative materialism: from folk psychology to Volkerpsychologie. *Philosophical Psychology* 5:349-68.

Haldane, J. 1988. Understanding folk. *Aristotelian Society Supplement* 62:222-46.

Argues that folk psychology is not a theory, and that psychological knowledge is a pre-theoretical given. With remarks on laws, the prediction of behavior, and neuroscience.

Hannan, B. 1990. 'Non-scientific realism' about propositional attitudes as a response to eliminativist arguments. *Behavior and Philosophy* 18:21-31.

Hannan, B. 1993. Don't stop believing: the case against eliminative materialism. *Mind and Language* 8:165-179.

A bundle of arguments against eliminativism, e.g. from incoherence, the lack of alternatives, and against the folk-theory-theory. With commentary.

Horgan, T. & Woodward, J. 1985. Folk psychology is here to stay. *Philosophical Review* 94:197-225. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Defending folk psychology against the arguments of Churchland and Stich: e.g. incompleteness, stagnation, irreducibility, dual-control, modularity, and unfalsifiability. Even with no neat reduction, folk psychology may be OK.

Horgan, T. & Graham, G. 1990. In defense of Southern Fundamentalism. *Philosophical Studies* 62:107-134.

FP is almost certainly true, irrespective of scientific absorbability or the language of thought. FP's commitments are austere, and mostly behavioral. Arguments from semantic competence and conceptual conservatism.

Horgan, T. 1993. The austere ideology of folk psychology. *Mind and Language*. Argues that FP is not committed to much. The austere conception is supported by intuitions, conservatism, and the inconceivability of dropping it. Responds to phlogiston objections: they are not analogous.

Horst, S. 1995. Eliminativism and the ambiguity of 'belief'. *Synthese* 104:123-45.

Clarifies different senses of "theoretical" and "belief". Some beliefs are relevantly theoretical (dispositional, infra-conscious, unconscious

ones), but conscious occurrent beliefs are not, and so can't be eliminated.

Jackson, F. & Pettit, P. 1990. In defense of folk psychology. *Philosophical Studies* 59:31-54.

FP holds that beliefs/desires play a certain functional role, and it's almost certain that objects playing that role exist, so FP is fine, whether or not propositional attitudes are good scientific entities.

Jacoby, H. 1985. Eliminativism, meaning and qualitative states. *Philosophical Studies*.

Even if nothing satisfies all or most common-sense properties of mental terms, reference can still be fixed under a Putnam style theory of meaning. (More about qualia than about intentional states.)

Kitcher, P.S. 1984. In defense of intentional psychology. *Journal of Philosophy* 81:89-106.

The Churchlands underestimate the resources of intentional psychology.

Lahav, R. 1992. The amazing predictive power of folk psychology. *Australasian Journal of Philosophy* 70:99-105.

Melnyk, A. 1996. Testament of a recovering eliminativist. *Philosophy of Science* 63:S185-93.

O'Brien, G. 1987. Eliminative materialism and our psychological self-knowledge. *Philosophical Studies* 52:49-70.

Uses empirical evidence to argue that there is prelinguistic awareness, so nominalistic arguments for eliminativism fail. And some awareness is innate, so we can't reconceive things in less than evolutionary time.

Ramsey, W. 1990. Where does the self-refutation objection take us? *Inquiry* 33:453-65.

The self-refutation objection reduces to other standard objections: counterexample, promissory note or reductio.

Ramsey, W., Stich, S.P. & Garon, J. 1991. Connectionism, eliminativism, and the future of folk psychology. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

If connectionism is true, then eliminativism is true, as you can't isolate the causal role of individual beliefs in a connectionist system.

Reppert, V. 1991. Ramsey on eliminativism and self-refutation. *Inquiry* 34:499-508.

Response to Ramsey 1990: If there are no beliefs and so no assertions, there is no identifiable propositional content, and truth and knowledge are out. Eliminativism is pragmatically self-refuting.

Reppert, V. 1992. Eliminative materialism, cognitive suicide, and begging the question. *Metaphilosophy* 23:378-92.

A careful analysis of whether self-refutation arguments against eliminativism beg the question by supposing that assertion requires belief. An account of what it is to beg the question, and a comparison to arguments about vitalism.

Resnick, P. 1994. Intentionality is phlogiston. In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

Richards, G. 1996. On the necessary survival of folk psychology. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Robinson, W.S. 1985. Toward eliminating Churchland's eliminationism. *Philosophical Topics* 13:60-67.

There's no reason to abandon FP, even if it doesn't reduce.

Rosenberg, A. 1991. How is eliminative materialism possible? In (R. Bogdan, ed) Mind and Common Sense. Cambridge University Press.
Explaining how singular causal claims based on FP may be true even if FP is false; by analogy with phlogiston, and also because of near-vacuousness. EM isn't incoherent, as we can use a non-intentional replacement for belief.

Rosenberg, A. 1999. Naturalistic epistemology for eliminative materialists. *Philosophy and Phenomenological Research* 59:335-358.

Saidel, E. 1992. What price neurophilosophy? *Philosophy of Science Association* 1:461-68.

Folk psychology is compatible with neuroscientific models, but it need not smoothly reduce to neuroscience to have an important role.

Schouten, M.K.D. & de Jong, H.L. 1998. Defusing eliminative materialism: Reference and revision. *Philosophical Psychology* 11:489-509.

Schwartz, J. 1991. Reduction, elimination, and the mental. *Philosophy of Science* 58:203-20.

Stich, S.P. 1991. Do true believers exist? *Aristotelian Society Supplement* 65:229-44.

Eliminativism may have no determinate truth-conditions: if folk psychology is a poor theory, the question of whether or not "belief" refers may be empty.

Stich, S.P. 1992. What is a theory of mental representation? *Mind* 101:243-61.
Philosophical analysis isn't sufficient to understand intentional concepts; real cognitive science is required, with conceptual revision. The truth of eliminativism will be relative to the theory of reference that we choose.

Stich, S.P. 1996. Deconstructing the mind. In Deconstructing the Mind. Oxford University Press, 1996.

Taylor, K.A. 1994. How not to refute eliminative materialism. *Philosophical Psychology* 7:101-125.

Against transcendental arguments against eliminativism. These fail on their own terms, and even if successful they would not establish causal/explanatory relevance for the attitudes, which is the real key for folk psychology.

Tomberlin, J. 1994. Whither Southern Fundamentalism? In (E. Villanueva, ed) Truth and Rationality, Ridgeview.

Trout, J.D. 1991. Belief attribution in science: Folk psychology under theoretical stress. *Synthese* 87:379-400.

Wright, C. 1996. Can there be a rationally compelling argument for anti-realism about ordinary ("folk") psychology? In (E. Villanueva, ed) Content. Ridgeview.

2.1d Propositional Attitudes, General

Audi, R. 1994. Dispositional beliefs and dispositions to believe. *Nous* 28:419-34.

Baker, L.R. 1987. Saving Belief. Princeton University Press.
Beliefs are OK, despite no physicalist reduction of content.

Baker, L.R. 1993. What beliefs are not. In (S. Wagner & R. Warner, eds) Naturalism: A Critical Appraisal. University of Notre Dame Press.
Against beliefs construed as physically realized internal causes of behavior: syntax of these states can't be determinate, and their explanatory role wrt

causation leads to a circle. Belief is irreducible.

Baker, L.R. 1994. Attitudes as nonentities. *Philosophical Studies* 76:175-203.

Balaguer, M. 1998. Attitudes without propositions. *Philosophy and phenomenological research* 58:805-26.

Bennett, J. 1991. Analysis without noise. In (R. Bogdan, ed) *_Mind and Common Sense_*. Cambridge University Press.

Remarks on the conceptual analysis of belief/desire attribution. On the roles of causation, inner-route explanations, belief-desire-action triangles, teleology, unity, the presumption of simplicity, and evolution.

Bennett, J. 1991. Folk-psychological explanations. In (J. Greenwood, ed) *_The Future of Folk Psychology_*. Cambridge University Press.

On requirements for belief/desire explanations: input/output patterns, the unity condition (i.e. no single associated mechanism), and teleological bases for generalizations, e.g. through evolution or educability.

Ben-Yami, H. 1997. Against characterizing mental states as propositional attitudes. *Philosophical Quarterly* 186:84-89.

Butler, K. 1992. The physiology of desire. *Journal of Mind and Behavior* 13:69-88.

Argues that desire will smoothly reduce to a neurophysiological kind.

Clark, A. 1991. Radical ascent. *Aristotelian Society Supplement* 65:211-27.

The conditions on being a believer are mostly behavioral; to claim otherwise is to fall into a "modularity trap". A counterfactual account of mental causation is enough. With a defense of mentality for giant look-up tables.

Clark, A. 1994. Beliefs and desires incorporated. *Journal of Philosophy* 91:404-25.

Cohen, L.J. 1996. Does belief exist? In (A. Clark & P. Millican, eds) *_Connectionism, Concepts, and Folk Psychology_*. Oxford University Press.

Crimmins, M. 1992. Tacitness and virtual beliefs. *Mind and Language* 7:240-63.

Davies, D. 1995. Davidson, indeterminacy, and measurement. *Acta Analytica* 10:37-56.

Davies, D. 1998. On gauging attitudes. *Philosophical Studies* 90:129-54.

Egan, M.F. 1989. What's wrong with the Syntactic Theory of Mind. *Philosophy of Science* 56:664-74.

Stich is confused about type-token, syntax/content, etc.

Fodor, J.A. 1986. Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind* 94:76-100. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

A taxonomy of positions on the representation of propositional attitudes: dividing up via questions about realism, functionalism, monadicity, and truth-conditions. With arguments for structured representations.

Frankish, K. 1998. A matter of opinion. *Philosophical Psychology* 11:423-442.

Garfield, J. 1988. *_Belief in Psychology: A Study in the Ontology of Mind_*. MIT Press.

Graham, G. & Horgan, T. 1988. How to be realistic about folk psychology. *Philosophical Psychology* 1.

- Jacquette, D. 1990. Intentionality and Stich's theory of brain sentence syntax. *Philosophical Quarterly*, 40:169-82.
Things are only syntactic (in SS's sense) in virtue of intentionality. True.
- Lycan, W.G. 1986. Tacit belief. In (R. Bogdan, ed) *_Belief: Form, Content, and Function_*. Oxford University Press.
- Maloney, J.C. 1990. It's hard to believe. *Mind and Language* 5:122-48.
- Manfredi, P.A. 1993. Tacit beliefs and other doxastic attitudes. *Philosophia*.
Argues that there are no tacit beliefs: dispositions to believe can do all the explanatory work at lower cost. With some remarks on subdoxastic states, and the difference between belief and opinion.
- Matthews, R.J. 1994. The measure of mind. *Mind* 103:131-46.
A theory of propositional attitude ascription as like numerical measurement.
- Millikan, R.G. 1986. Thoughts without laws: Cognitive science with content. *Philosophical Review* 95:47-80.
Folk psychology isn't a theory about laws, but about proper functions.
desires are identified by proper functions; beliefs by Normal explanations.
- Moser, P.K. 1990. Physicalism and intentional attitudes. *Behavior and Philosophy* 18:33-41.
- Peacocke, C. 1983. Between instrumentalism and brain-writing. In *_Sense and Content_*. Oxford University Press.
Instrumentalism about belief can't be right, because of Martian marionettes, but the language of thought is too strong a requirement. A state's structured content may reside in its pattern of relations to other states.
- Possin, K. 1986. The case against Stich's Syntactic Theory of Mind. *Philosophical Studies* 49:405-18.
Stich is wrong, circular, and representational anyway.
- Pratt, I. 1993. Analysis and the attitudes. In (S. Wagner & R. Warner, eds) *_Naturalism: A Critical Appraisal_*. University of Notre Dame Press.
- Pylyshyn, Z.W. 1987. What's in a mind? *Synthese* 70:97-122.
Must individuate mental states by semantics, not just by function.
- Recanati, F. 1997. Can we believe what we do not understand? *Mind and Language* 12:84-100.
- Robinson, W.S. 1990. States and beliefs. *Mind* 99:33-51.
- Schwartz, J. 1992. Propositional attitude psychology as an ideal type. *Topoi* 11:5-26.
- Smith, D.M. 1994. Toward a perspicuous characterization of intentional states. *Philosophical Studies* 74:103-20.
- Sobel, D. & Copp, D. 2001. Against direction of fit accounts of belief and desire. *Analysis* 61:44-53.
- Sperber, D. 1997. Intuitive and reflective beliefs. *Mind and Language* 12:67-83.
- Stich, S.P. 1983. *_From Folk Psychology to Cognitive Science_*. MIT Press.
Beliefs/desires are out, new Syntactic Theory is in.
- Stich, S.P. 1984. Relativism, rationality, and the limits of intentional ascription. *Pacific Philosophical Quarterly*.

Stone, T. & Young, A.W. 1997. Delusions and brain injury: The philosophy and psychology of belief. *Mind and Language* 12:327-364.

Von Eckardt, B. & Poland, J. 2000. In defense of the standard view. *Protosociology* 14:312-331.

Weatherall, P. 1996. What do propositions measure in folk psychology? *Philosophical Psychology* 9:365-80.

2.1e The Nature of Folk Psychology

Blackburn, S. 1992. Theory, observation, and drama. *Mind and Language* 7:187.

Bogdan, R.G. (ed) 1991. *Mind and Common Sense: Philosophical Essays on Commonsense Psychology*. Cambridge University Press.

Botterill, G. 1996. Folk psychology and theoretical status. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Churchland, P.M. 1988. Folk psychology and the explanation of human behavior. *Proceedings of the Aristotelian Society* 62:209-21. Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Folk psychology is a theory: defense against objections from logicality, softness of laws, practical function, behavior, and simulation. It needn't be a deductive-nomological theory; e.g. it might be based on prototypes.

Clark, A. 1987. From folk psychology to naive psychology. *Cognitive Science* 11:139-54.

Folk psychology isn't all that bad. It survived evolution after all.

Collins, J. 2000. Theory of mind, logical form and eliminativism. *Philosophical Psychology* 13:465-490.

Dennett, D.C. 1991. Two contrasts: Folk craft vs folk science and belief vs opinion. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

FP is craft, not theory. Opinions rather than beliefs are interesting.

Falvey, K. 1999. A natural history of belief. *Pacific Philosophical Quarterly* 80:324-345.

Fletcher, G. 1995. *The Scientific Credibility of Folk Psychology*. Lawrence Erlbaum.

Fletcher, G. 1995. Two uses of folk psychology: Implications for psychological science. *Philosophical Psychology* 8:375-88.

Goldman, A. 1992. The psychology of folk psychology. *Behavioral and Brain Sciences*.

On the psychology of self-ascription of mental states. Functionalism has serious problems, as we don't have direct access to causal roles. Defends a qualia-based account, even for propositional attitudes.

Gopnik, A. 1990. Developing the idea of intentionality: Children's theories of mind. *Canadian Journal of Philosophy* 20:89-114.

On the development of folk-psychological concepts in children. First the appearance/reality distinction, then more complex theories of perception, representation, and belief. Implications for the status of folk psychology.

Gopnik, A. & Wellman, H. 1992. Why the child's theory of mind really is a theory. *Mind and Language* 7:145-71.

- Graham, G. 1987. The origins of folk psychology. *Inquiry* 30:357-79.
- Greenwood, J.D. (ed) 1991. *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge University Press.
- Leon, M. 1998. The unnaturalness of the mental: The status of folk psychology. *Southern Journal of Philosophy* 36:367-92.
- Lycan, W.G. 1997. Folk psychology and its liabilities. In (M. Carrier & P. Machamer, eds) *Mindsapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.
- Margolis, J. 1991. The autonomy of folk psychology. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.
- McDonough, R. 1991. A culturalist account of folk psychology. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.
- Morton, A. 1980. *Frames of Mind*. Oxford University Press.
- Morton, A. 1991. The inevitability of folk psychology. In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.
- Morton, A. 1996. Folk psychology is not a predictive device. *Mind* 105:119-37.
- Pettit, P. 2000. How the folk understand folk psychology. *Protosociology* 14:26-38.
- Place, U.T. 1996. Folk psychology from the standpoint of conceptual analysis. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.
- Pratt, I. 1996. Encoding psychological knowledge. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.
- Preston, J.M. 1989. Folk psychology as theory or practice? The case for eliminative materialism. *Inquiry* 32:277-303.
Defending the claim that folk psychology is an empirical pre-scientific theory, with its own laws. In a particular, a detailed reply to the criticisms in Wilkes 1984.
- Robinson, W.S. 1996. Mild realism, causation, and folk psychology. *Philosophical Psychology* 8:167-87.
- Schwitzgebel, E. 2001. In-between believing. *Philosophical Quarterly* 51:76 - 82.
- Sehon S.R. 1997. Natural kind terms and the status of folk psychology. *American Philosophical Quarterly* 34:333-44.
- Sharpe, R. 1987. The very idea of a folk psychology. *Inquiry* 30:381-93.
- Smith, B.C. 1996. Does science underwrite our folk psychology? In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.
- Stemmer, N. 1995. A behaviorist account to theory and simulatyion theories of folk psychology. *Behavior and Philosophy* 23:29-41.
- Sterelny, K. 1998. Intentional agency and the metarepresentation hypothesis. *Mind and Language* 13:11-28.

Stich, S.P. & Ravenscroft, R. 1994. What is folk psychology? *Cognition* 50:447-68. Reprinted in (Stich) *_Deconstructing the Mind_*. Oxford University Press, 1996.

Distinguishes internal and external accounts of folk psychology (mechanisms vs systematizations), and various versions of each of these. Only some are compatible with eliminativist arguments.

von Eckardt, B. 1997. The empirical naivete in the current philosophical conception of folk psychology. In (M. Carrier & P. Machamer, eds) *_Mindsapes: Philosophy, Science, and the Mind_*. Pittsburgh University Press.

Wilkes, K.V. 1984. Pragmatics in science and theory in common sense. *Inquiry* 27:339-61.

Wilkes, K.V. 1991. The relationship between scientific psychology and common-sense psychology. *Synthese* 89:15-39.

Common-sense psychology is no theory at all, and not in competition with scientific psychology. CSP is particular, rich, vague; SP is general, austere, precise. CSP will be neither subsumed nor eliminated by SP.

Wilkes, K.V. 1991. The long past and the short history. In (R. Bogdan, ed) *_Mind and Common Sense_*. Cambridge University Press.

Argues that commonsense and scientific psychology are quite distinct in their aims, scope, framework, and nature, but have been confused by philosophy. With support from historical considerations.

2.1f The Simulation Theory

Arkway, A. 2000. The simulation theory, the theory theory and folk psychological explanation. *Philosophical Studies* 98:115-137.

Carruthers, P. 1996. Simulation and self-knowledge: A defence of the theory-theory. In (P. Carruthers & P. Smith, eds) *_Theories of Theories of Mind_*. Cambridge University Press.

Carruthers, P. & Smith, P. 1996. *_Theories of Theories of Mind_*. Cambridge University Press.

Cruz, J.L.H. 1998. Mindreading: Mental state ascription and cognitive architecture. *Mind and Language* 13:323-340.

Currie, G. 1995. Visual imagery as the simulation of vision. *Mind and Language* 10:25-44.

Currie, G. 1996. Simulation-theory, theory-theory, and the evidence from autism. In (P. Carruthers & P. Smith, eds) *_Theories of Theories of Mind_*. Cambridge University Press.

Currie, G. & Ravenscroft, I. 1997. Mental simulation and motor imagery. *Philosophy of Science* 64:161-80/

Currie, G. 1998. Pretence, pretending, and metarepresenting. *Mind and Language* 13:35-55.

Davies, M. 1992. The mental simulation debate. In (E. Villanueva, ed) *_Truth and Rationality_*. Ridgeview.

Davies, M. & Stone, T. (eds) 1995. *_Folk Psychology: The Theory of Mind Debate_*. Blackwell.

Davies, M. & Stone, T. (eds) 1995. *_Mental Simulation: Evaluations and Applications_*. Blackwell.

Freeman, N.H. 1995. Theories of mind in collision: Plausibility and authority. In (M. Davies & T. Stone, eds) Mental Simulation. Blackwell.

Fuller, G. 1995. Simulation and psychological concepts. In (M. Davies & T. Stone, eds) Mental Simulation. Blackwell.

Goldman, A. 1989. Interpretation psychologized. *Mind and Language* 4:161-85. Reprinted in (M. Davies & T. Stone, eds) Folk Psychology. Blackwell.

Goldman, A. 1992. In defense of the simulation theory. *Mind and Language* 7:104-119. Reprinted in (M. Davies & T. Stone, eds) Folk Psychology. Blackwell.

Goldman, A. 1996. Simulation and interpersonal utility. In (L. May, M. Friedman, & A. Clark, eds) Mind and Morals: Essays on Ethics and Cognitive Science. MIT Press.

Goldman, A. 2000. Folk psychology and mental concepts. *Protosociology* 14:4-25.

Gopnik, A. & Wellman, H.M. 1995. Why the child's theory of mind really is a theory. *Mind and Language*. Reprinted in (M. Davies & T. Stone, eds) Folk Psychology. Blackwell.

Gopnik, A. and Meltzoff, AN. 1998. Theories vs. modules: To the max and beyond. A reply to Poulin-Dubois and to Stich and Nichols. *Mind and Language* 13:450-456.

Gordon, R.M. 1986. Folk psychology as simulation. *Mind and Language* 1:158-71. Reprinted in (M. Davies & T. Stone, eds) Folk Psychology. Blackwell.
FP is a strategy for prediction via simulation; an ability, not a theory.

Gordon, R.M. 1992. The simulation theory: objections and misconceptions. *Mind and Language* 7:11-34. Reprinted in (M. Davies & T. Stone, eds) Folk Psychology. Blackwell.

Gordon, R.M. & Barker, J.A. 1994. Autism and the "theory of mind" debate. In (G. Graham & G.L. Stephens, eds) Philosophical Psychopathology. MIT Press.

Gordon, R.M. 1995. Simulation without introspection or inference from me to you. In (M. Davies & T. Stone, eds) Mental Simulation. Blackwell.

Gordon, R.M. 1996. Sympathy, simulation, and the impartial spectator. In (L. May, M. Friedman, & A. Clark, eds) Mind and Morals: Essays on Ethics and Cognitive Science. MIT Press.

Gordon, R.M. 1996. 'Radical' simulationism. In (P. Carruthers & P. Smith, eds) Theories of Theories of Mind. Cambridge University Press.

Gordon, R.M. 2000. Sellars's Rylean ancestors revisited. *Protosociology* 14:102-114.

Greenwood, J.D. 1999. Simulation, theory-theory and cognitive penetration: No "instance of the fingerpost". *Mind and Language* 14:32-56.

Heal, J. 1986. Replication and functionalism. In (J. Butterfield, ed) Language, Mind, and Logic. Cambridge University Press. Reprinted in (M. Davies & T. Stone, eds) Folk Psychology. Blackwell.

Heal, J. 1994. Simulation vs. theory-theory: What is at issue? In (C. Peacocke, ed) Objectivity, Simulation, and the Unity of Consciousness. Oxford University Press.

Heal, J. 1995. How to think about thinking. In (M. Davies & T. Stone, eds)

Mental Simulation. Blackwell.

Heal, J. 1996. Simulation and cognitive penetrability. *Mind and Language* 11:44-67.

Heal, J. 1996. Simulation, theory, and content. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Heal, J. 1998. Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind and Language* 13:477-498.

Heal, J. 2000. Understanding other minds from the inside. *Protosociology* 14:39-55.

Henderson, D. 1996. Simulation theory versus theory theory: A difference without a difference in explanations. *Southern Journal of Philosophy* 34:65-93.

Kuhberger, A, Perner, J., Schulte, M., & Leingruber, R. 1995. Choice or no choice: Is the Langer effect evidence against simulation? *Mind and Language* 10:423-36.

Leslie, A.M. & German, T.P. 1995. Knowledge and ability in "theory of mind": A one-eyed overview of a debate. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Levin, J. 1995. Folk psychology and the simulationist challenge. *Acta Analytica* 10:77-100.

Nichols, S., Stich, S., & Leslie, A. 1995. Choice effects and the ineffectiveness of simulation. *Mind and Language* 10:437-45.

Nichols, S., Stich, S., Leslie, A., Klein, D. 1996. Varieties of off-line simulation. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Nichols, S. & Stich, S. 1998. Rethinking co-cognition: A reply to Heal. *Mind and Language* 13:499-512.

Perner, J. 1994. The necessity and impossibility of simulation. In (C. Peacocke, ed) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford University Press.

Perner, J. 1996. Simulation as explicitation of predication-implicit knowledge about the mind: Arguments for a simulation-theory mix. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Perner, J., Gschaidner, A., Kuhberger, A. & Schrofner, S. 1999. Predicting others through simulation or by theory? A method to decide. *Mind and Language* 14:57-79.

Pust, J. 1999. External accounts of folk psychology, eliminativism, and the simulation theory. *Mind and Language* 14:113-130.

Ruffman, T. 1996. Do children understand the mind by means of a simulation or a theory? Evidence from their understanding of inference. *Mind and Language* 11:388-414.

Scholl, B.J. & Leslie, A.M. 1999. Modularity, development and "theory of mind." *Mind and Language* 14:131-153.

Schwitzgebel, E. 1999. Representation and desire: a philosophical error with consequences for theory-of-mind research. *Philosophical Psychology* 12:157-180.

Stich, S.P. & Nichols, S. 1993. Folk psychology: simulation or tacit theory? *Mind and Language* 7:35-71. Reprinted in (M. Davies & T. Stone, eds) *Folk Psychology*. Blackwell.

Stich, S.P. & Nichols, S. 1995. Second thoughts on simulation. In (M. Davies & T. Stone, eds) *Mental Simulation*. Blackwell.

Stich, S.P. & Nichols, S. 1997. Cognitive penetrability, rationality, and restricted simulation. *Mind and Language* 12:297-326.

Stich, S. & Nichols, S. 1998. Theory theory to the max. *Mind and Language* 13:421-449.

Stone, T. & Davies, M. 1996. The mental simulation debate: A progress report. In (P. Carruthers & P. Smith, eds) *Theories of Theories of Mind*. Cambridge University Press.

Wilkerson, W.S. 2001. Simulation, theory, and the frame problem: the interpretive moment. *Philosophical Psychology* 14:141-153.

2.2 Internalism and Externalism [see also 1.5d]

2.2a Is Content in the Head? (Putnam, Burge)

Antony, M. 1993. Social relations and the individuation of thought. *Mind* 102:247-61.

Bilgrami, A. 1987. An externalist account of psychological content. *Philosophical Topics* 15:191-226.

Developing an externalist account consistent with psychological explanation.
Contra Burge, social links aren't constitutive of content. Causal links
are indirectly constitutive of content, via our conceptions.

Brueckner, A. 1995. The characteristic thesis of anti-individualism. *Analysis* 55:146-48.

Bruns, M. & Soldati, G. 1997. Object-dependent and property-dependent concepts. *Dialectica* 48:185-208.

Burge, T. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4:73-122.

Belief contents are not fully determined by internal state, as the linguistic
community plays an important role: arthritis, brisket, contract, sofa, etc.
So mental states are not individuated individualistically.

Burge, T. 1982. Other bodies. In (A. Woodfield, ed) *Thought and Object*.
Oxford University Press.

On Putnam's Twin Earth. Natural kind terms are not indexical. Even de dicto
attitudes are not in the head; they presuppose the existence of other things.

Burge, T. 1986. Intellectual norms and foundations of mind. *Journal of Philosophy* 83:697-720.

On non-individualist elements due to by intellectual norms in the community,
to which meanings are answerable. Even meaning-giving truths can be doubted.
With remarks on sofas/safos, and on linguistic meaning vs. cognitive value.

Butler, K. 1993. Individualism, computationalism, and folk psychology.
Manuscript.

Challenges Burge's interpretations of the thought-experiments: e.g. twins
have the same concept, neither of which is the public concept of arthritis.
With remarks on computationalism and Marr's theory.

- Campbell, J. 1982. Extension and psychic state: Twin Earth revisited. *Philosophical Studies* 42:67-89.
Argues that natural kind terms are token-reflexive, with reference ultimately fixed to the underlying explanatory properties of the surface qualities of local matter.
- Crane, T. 1991. All the difference in the world. *Philosophical Quarterly* 41:1-25.
Twins share the same concepts. Contra Putnam: essentialism is fallacious; contra Burge: speakers share beliefs, but one has false belief about meaning.
- Cummins, R. 1991. Methodological reflections on belief. In (R. Bogdan, ed) *Mind and Common Sense*. Cambridge University Press.
We shouldn't rely on intuitions about thought-experiments; we need an empirical theory about belief. Belief contents are distinct from sentence contents; we have to distinguish linguistic from psychological semantics.
- Devitt, M. 1990. Meanings just ain't in the head. In (G. Boolos, ed) *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.
Against Searle's theory of internal intentionality. Searle's theory requires magic to grasp external contents internally.
- Dretske, F. 1993. The nature of thought. *Philosophical Studies* 70:185-99.
Argues that thought is extrinsic, but it is not essentially social. A system without a linguistic community could have thoughts, if it had an appropriate learning history.
- Elugardo, R. 1993. Burge on content. *Philosophy and Phenomenological Research* 53:367-84.
Contra Burge on sofas: oblique that-clauses can't identify the (wide) way that the subject thinks of sofas, which is idiosyncratic and inexpressible.
- Forbes, G. 1987. A dichotomy sustained. *Philosophical Studies* 51:187-211.
Gives a Fregean account of belief semantics to handle the Burge cases, and argues that the *type* of a proposition may be internal even if the token itself is not. With remarks on the relevance to Grice's program.
- Georgalis, N. 1999. Rethinking Burge's thought experiment. *Synthese* 118:145-64.
- Horowitz, A. 1995. Putnam, Searle, and externalism. *Philosophical Studies* 81:27-69.
Argues for a moderate externalism by synthesizing Putnam and Searle: internal intension leaves extension indeterminate, but it specifies the facts relevant to filling in the indeterminacy.
- Koethe, J. 1992. And they ain't outside the head either. *Synthese* 90:27-53.
- Ludwig, K. 1993. Externalism, naturalism, and method. In (E. Villanueva, ed) *Naturalism and Normativity*. Ridgeview.
- Ludwig, K. 1996. Duplicating thoughts. *Mind and Language* 11:92-102.
- Mandelkar, S. 1991. An argument against the externalist account of psychological content. *Philosophical Psychology* 4:375-82.
Argues that conscious experience is required for intentional states, and that any external relations could be satisfied without this experience, so external relations cannot suffice for intentional content.
- McCulloch, G. 1992. The spirit of twin earth. *Analysis* 52:168-174.
Various arguments against Crane 1991 on externalism.

McDowell, J. 1977. On the sense and reference of a proper name. *Mind*.

McGilvray, J. 1998. Meanings are syntactically individuated and found in the head. *Mind and Language* 13:225-280.

McKinsey, M. 1991. The internal basis of meaning. *Pacific Philosophical Quarterly* 72:143-69.

Argues that meaning is determined by a certain kind of internal state, involving de se cognitive attitudes. These states aren't shared by twins, but are still narrow in a strong sense.

McKinsey, M. 1993. Curing folk psychology of arthritis. *Philosophical Studies* 70:323-36.

McKinsey, M. 1994. Individuating beliefs. *Philosophical Perspectives* 8:303-30.

Owens, J. 1983. Functionalism and the propositional attitudes. *Nous* 17:529-49.

Functional organization doesn't determine attitude content, even if we include inputs and outputs.

Perry, J. 1979. The problem of the essential indexical. *Nous* 13:3-21.

Indexicals are essential to some beliefs, so belief cannot just be a relation to a proposition. Belief contents must be at least in part construed relative to a subject. Separate belief object and belief state.

Putnam, H. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7:131-193. Reprinted in *_Mind, Language, and Reality_* (Cambridge University Press, 1975).

What is in the head doesn't determine the reference of our thoughts: my twin on Twin Earth refers to XYZ where I refer to H₂O. Content is determined by environment and linguistic community as well as by internal stereotypes.

Putnam, H. 1987. Meaning, other people, and the world. In *_Representation and Reality_*. MIT Press.

Meanings **still** aren't in the head.

Searle, J.R. 1983. *_Intentionality_*. Cambridge University Press.

Sure, meanings **are** in the head -- e.g. the content of a given visual experience is "the thing that is causing this experience".

Sosa, E. 1991. Between internalism and externalism. In (E. Villanueva, ed) *_Consciousness_*. Ridgeview.

Sosa, E. 1993. Abilities, concepts, and externalism. In (J. Heil & A. Mele, eds) *_Mental Causation_*. Oxford University Press.

On concepts as abilities, and on construals of abilities that lead to internalism and externalism. Maybe the relevant abilities are characterized externally but determined internally. Remarks on Putnam, Davidson, Burge.

Stalnaker, R. 1993. Twin earth revisited. *Proceedings of the Aristotelian Society* 63:297-311.

Making sense of twin earth intuitions with an information-theoretic account of content: information depends on relations in normal conditions, which are extrinsic. With remarks on the context-sensitivity of content-attribution.

Wikforss, A. 2001. Social externalism and conceptual errors. *Philosophical Quarterly* 203:217-31.

Woodfield, A. 1982. Thought and the social community. *Inquiry* 25:435-50.

Burge's arguments show only that context-ascription is pragmatically sensitive to context, depending on the epistemic predicament of the ascriber.

Content itself is still internal.

Zemach, E.M. 1976. Putnam's theory on the reference of substance terms.
Journal of Philosophy 73:116-27.

Argues that the extension of 'water' is the same on earth and twin earth,
using arguments from isotopes and scientific development. Molar properties
determine classification. Remarks on historicism and the division of labor.

2.2b Externalism and Psychological Explanation (Burge, Fodor)

Adams, F. & K. Aizawa, 2001. The bounds of cognition. Philosophical
Psychology 14:43-64.

Arjo, D. 1996. Sticking up for Oedipus: Fodor on intentional generalizations
and broad content. Mind and Language 11:231-45.

Buller, D.J. 1992. "Narrow"-minded breeds inaction. Behavior and Philosophy
20:59-70.

Buller, D.J. 1997. Individualism and evolutionary psychology (or: In defense
of "narrow" functions). Philosophy of Science 64:74-95.

Burge, T. 1982. Two thought experiments reviewed. Notre Dame Journal of
Formal Logic 23:284-94.

Reply to Fodor 1982, clarification of position.

Burge, T. 1986. Individualism and psychology. Philosophical Review 95:3-45.
Psychology should be and is done non-individualistically, i.e. with reference
to environment. Examples from vision, e.g. Marr.

Clark, A. & Chalmers, D.J. 1998. The extended mind. Analysis 58:7-19.
Advocates a different sort of "active externalism", based on the role of the
environment in actively driving cognition. Beliefs can extend into an
agent's immediate environment (e.g. a notebook) in this way.

Dretske, F. 1992. What isn't wrong with folk psychology. Metaphilosophy
23:1-13.

Argues that extrinsic properties can play a respectable role in scientific
explanation; e.g. the histories of plants, animals, and devices are relevant
in explaining their current behavior.

Egan, F. 1991. Must psychology be individualistic? Philosophical Review
100:179-203.

Maybe, maybe not. Contra Fodor: science can be non-individualistic.

Contra Burge re oblique ascriptions and Marr.

Fodor, J.A. 1980. Methodological solipsism as a research strategy in cognitive
psychology. Behavioral and Brain Sciences 3:63-109. Reprinted in
RePresentations (MIT Press, 1980).

Should do psychology without reference to the external world. What counts for
psychology is in the head; who cares about truth, reference, and the rest?

Fodor, J.A. 1982. Cognitive science and the twin-earth problem. Notre Dame
Journal of Formal Logic 23:98-118.

Twin Earth isn't a problem for cognitive science. Intents of utterances, de
re/de dicto, etc. Truth conditions aren't in the head, but that's no problem.

Gauker, C. 1987. Mind and chance. Canadian Journal of Philosophy 17:533-52.

Globus, G. 1984. Can methodological solipsism be confined to psychology?
Cognition and Brain Theory 7:233-46.

Methodological solipsism implies epistemological solipsism.

Hardcastle, V.G. 1997. [Explanation] is explanation better. *Philosophy of Science* 64:154-60.

Hurley, S.L. 1998. Vehicles, contents, conceptual structure, and externalism. *Analysis* 58:1-6.

Jacob, P. 1993. Externalism and the explanatory relevance of broad content. *Mind and Language* 8:131.

Kitcher, P.S. 1984. Narrow taxonomy and wide functionalism. *Philosophy of Science* 52:78-97.

Argues against Stich, Fodor, Block: use different taxonomies (narrow/wide) for different purposes. Both are OK, functionalism *and* content survive.

Kobes, B. 1989. Semantics and psychological prototypes. *Pacific Philosophical Quarterly* 70:1-18.

Relates the individualism debate to Roschian prototype research.

Losonsky, M. 1995. Emdedded systems vs. individualism. *Minds and Machines* 5:357-71.

Macdonald, C. 1992. Weak externalism and psychological reduction. In (D Charles & K. Lennon, eds) *_Reduction, Explanation and Realism_*. Oxford University Press.

Marras, A. 1985. The Churchlands on methodological solipsism and computational psychology. *Philosophy of Science* 52:295-309.

MS doesn't rule out all use of content, just of wide content. Narrow content is OK. With remarks on folk psychology and computation.

Maloney, J.C. 1985. Methodological solipsism reconsidered as a research strategy in cognitive psychology. *Philosophy of Science* 52:451-69.

Various problems for computational psychology handling content. It shares the problems of a naturalistic psychology.

McClamrock, R. 1991. Methodological individualism considered as a constitutive principle of scientific inquiry. *Philosophical Psychology* 4:343-54.

McClamrock, R. 1995. *_Existential Cognition: Computational Minds in the World_*. University of Chicago Press.

Noonan, H.W. 1984. Methodological solipsism: A reply to Morris. *Philosophical Studies* 48:285-290.

Noonan, H.W. 1986. Russellian thoughts and methodological solipsism. In (J. Butterfield, ed) *_Language, Mind, and Logic_*. Cambridge University Press.

Noonan, H.W. 1990. Object-dependent thoughts and psychological redundancy. *Analysis* 51:1-9.

Noonan, H.W. 1993. Object-dependent thoughts: A case of superficial necessity but deep contingency? In (J. Heil & A. Mele, eds) *_Mental Causation_*. Oxford University Press.

Object-dependent thoughts are redundant in psychological explanation, as an explanation applying to a hallucinator will work as well. But this needn't defeat externalism in general. With remarks on self-knowledge.

Patterson, S. 1990. The explanatory role of belief ascriptions. *Philosophical Studies* 59:313-32.

Uses examples to argue that in explaining behavior we often ascribe beliefs in an individualistic way, even in cases where individual and community use diverge. These contents are at least sometimes expressible.

Patterson, S. 1991. Individualism and semantic development. *Philosophy of Science* 58:15-35.

Developmental psychologists attribute concepts individualistically.

Peacocke, C. 1993. Externalist explanation. *Proceedings of the Aristotelian Society* 67:203-30.

Externalist states are required for the explanation of relational properties. Counters objections from conceptual connections and dormitive-virtue worries, and applies to teleology, self-knowledge, etc.

Petrie, B. 1990. Nonautonomous psychology. *Southern Journal of Philosophy* 28:539-59.

Argues that behavior is often individuated widely for explanatory purposes, so that wide content is relevant, and that there is more to causation than local causation, so Stich's autonomy principle fails.

Pettit, P. 1986. Broad-minded explanation and psychology. In (P. Pettit & J. McDowell, eds) *_Subject, Thought and Context_*. Oxford University Press.

Rowlands, M. 1995. Against methodological solipsism: The ecological Approach. *Philosophical Psychology* 8:5-24.

Segal, G. 1989. The return of the individual. *Mind* 98:39-57.

Sterelny, K. 1990. Animals and individualism. In (P. Hanson, ed) *_Information, Language and Cognition_*. University of British Columbia Press.

Stich, S.P. 1978. Autonomous psychology and the belief/desire thesis. *Monist* 61:573-91. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Beliefs are not in the head, so aren't good for psychological explanation. Interesting, but confuses the role of truth-values with truth-conditions.

Tuomela, R. 1989. Methodological solipsism and explanation in psychology. *Philosophy of Science* 56:23-47.

Wallace, J. & Mason, H.E. 1990. On some thought experiments about mind and meaning. In (C. Anderson & J. Owens, eds) *_Propositional Attitudes_*. CSLI.

Wilson, R.A. 1994. Causal depth, theoretical appropriateness, and individualism in psychology. *Philosophy of Science* 61:55-75.

Wilson, R.A. 1995. *_Cartesian Psychology and Physical Minds: Individualism and the Sciences of the Mind_*. Cambridge University Press.

2.2c Externalism and Mental Causation

Adams, F. 1993. Fodor's modal argument. *Philosophical Psychology* 6:41-56.

Allen, C. 1995. It isn't what you think: A new idea about intentional causation. *Nous* 29:115-26.

Baker, L.R. 1994. Content and context. *Philosophical Perspectives* 8:17-32.

Argues contra Fodor that broad contents can be explanatory -- if they can't, no relational properties can. Fodor's "no-conceptual-connection" and "cross-context" tests for causal powers fail to do the job.

Barrett, J. 1997. Individualism and the cross-contexts test. *Pacific Philosophical Quarterly* 78:242-??.

Braun, D. 1991. Content, causation, and cognitive science. *Australasian Journal of Philosophy* 69:375-89.

Arguments for the causal significance of broad content. Physical twins can

differ in causal powers; broad content figures in (*ceteris paribus*) causal generalizations; can make twin arguments against narrow content too. Hmm.

Burge, T. 1989. Individuation and causation in psychology. *Pacific Philosophical Quarterly* 707:303-22.

Contra Fodor: psychological processes can play differing causal roles, despite being physically identical.

Burge, T. 1995. Intentional properties and causation. In (C. Macdonald & G. Macdonald, eds) *Philosophy of Psychology: Debates about Psychological Explanation*. Blackwell.

Reply to Fodor 1991.

Butler, K. 1996. Content, causal powers, and context. *Philosophy of Science* 63:105-14.

Christensen, D. 1992. Causal powers and conceptual connections. *Analysis* 52:163-8.

Fodor's modal argument for narrow content rests on a false analogy between cases concerning thoughts and those concerning planets.

Fodor, J.A. 1991. A modal argument for narrow content. *Journal of Philosophy* 88:5-26.

On when a difference in effects amounts to a difference in causal powers: when the effects are connected contingently, not conceptually, to the causes. Differences in wide content don't satisfy this, so aren't causal powers.

Garcia-Carpintero, M. 1994. The supervenience of mental content. *Proceedings of the Aristotelian Society* 68:117-135.

Mental content can be extrinsic and efficacious. Narrow content strategies don't work, as observation concepts are still extrinsic. One can't screen off the intrinsic part from the rest. Thought-experiments are inconclusive.

Heil, J. & Mele, A. 1991. Mental causes. *American Philosophical Quarterly* 28:61-71.

Reconciling Twin Earth with the causal relevance of content. Historical factors can be causally relevant.

Jacob, P. 1992. Externalism and mental causation. *Proceedings of the Aristotelian Society* 66:203-19.

Argues that externalist content is not causally efficacious, but is relevant to causal explanations of behavior indirectly, via the cognitive activities of others external to the system.

Klein, M. 1996. Externalism, content, and causation. *Proceedings of the Aristotelian Society* 96:159-76.

Lalor, B.J. 1997. It is what you think: intentional potency and anti-individualism. *Philosophical Psychology* 10:165-78.

Ludwig, K. 1993. Causal relevance and thought content. *Philosophical Quarterly* 44:334-53.

McGinn, C. 1991. Conceptual causation. *Mind* 100:525-46.

Montgomery, R. 1995. Non-Cartesian explanations meet the problem of mental causation. *Southern Journal of Philosophy* 33:221-41.

Owens, J. 1993. Content, causation, and psychophysical supervenience. *Philosophy of Science* 60:242-61.

Russow, L.M. 1993. Fodor, Adams, and causal properties. *Philosophical Psychology* 6:57-61.

Saidel, E. 1994. Content and causal powers. *Philosophy of Science* 61:658-65.

Segal, G. & Sober, E. 1991. The causal efficacy of content. *Philosophical Studies* 63:1-30.

Seymour, D. 1993. Some of the difference in the world: Crane on intentional causation. *Philosophical Quarterly* 43:83-89.

Sturgeon, S. 1994. Good reasoning and cognitive architecture. *Mind and Language* 9:88-101.

Epistemology requires the causal relevance of content, and the relevant content is narrow. On how various architectures might support this causal relevance, by being realized by more specific intrinsic features.

van Gulick, R. 1989. Metaphysical arguments for internalism and why they don't work. In (S. Silvers, ed) *_ReRepresentation_*. Kluwer.

Against some arguments for internalism: local causation doesn't imply local type-individuation, as distal relations affect distal causes and effects; and processes can have access to semantic properties via formal properties.

Wilson, R.A. 1992. Individualism, causal powers, and explanation. *Philosophical Studies* 68:103-39.

Science frequently appeals to relational and historical taxonomies, so either causal powers can be non-intrinsic or science needn't taxonomize by causal powers. With remarks on causal properties and conceptual connections.

Wilson, R.A. 1993. Against a priori arguments for individualism. *Pacific Philosophical Quarterly* 74:60-79.

Arguments from causal powers beg the question, either on whether relational properties can have causal powers or on whether science taxonomizes by causal powers, as relational properties are common in scientific explanation.

Yablo, S. 1997. Wide causation. *Philosophical Perspectives* 11:251-81.

2.2d Externalism and the Theory of Vision

Burge, T. 1986. Individualism and psychology. *Philosophical Review* 95:3-45.

Psychology should be and is done non-individualistically, i.e. with reference to environment. Examples from vision, e.g. Marr.

Butler, K. 1996. Individualism and Marr's computational theory of vision. *Mind and Language* 11:313-37.

Butler, K. 1996. Content, computation, and individualism in vision theory. *Analysis* 56:146-54.

Cain, M.J. 2000. Individualism, twin scenarios and visual content. *Philosophical Psychology* 13:441-463.

Davies, M. 1991. Individualism and perceptual content. *Mind* 100:461-84.

Egan, F. 1992. Individualism, computation, and perceptual content. *Mind* 101:443-59.

Egan, F. 1996. Intentionality and the theory of vision. In (K. Akers, ed) *_Perception_*. Oxford University Press.

Francescotti, R.M. 1991. Externalism and the Marr theory of vision. *British Journal for the Philosophy of Science* 42:227-38.

Kitcher, P.S. 1988. Marr's computational theory of vision. *Philosophy of*

Science 55:1-24.

Morton, P. 1993. Supervenience and computational explanation in vision theory. *Philosophy of Science* 60:86-99.

Patterson, S. 1996. Success-orientation and individualism in the theory of vision. In (K. Akins, ed) *_Perception_*. Oxford University Press.

Segal, G. 1989. Seeing what is not there. *Philosophical Review* 97:189-214.
Contra Burge, Marr's theory is individualistic. Intentional contents therein are neutral between twins' environments; nothing grounds a more specific attribution.

Segal, G. 1991. Defence of a reasonable individualism. *Mind* 100:485-94.

Shapiro, L.A. 1993. Content, kinds, and individualism in Marr's theory of vision. *Philosophical Review* 102:489-513.
Contra Segal, Marr's theory is non-individualistic even though it may classify twins together. Computational-level task descriptions rather than behavior guide content ascription, so the environment plays a crucial role.

Shapiro, L.A. 1997. A clearer vision. *Philosophy of Science* 64:131-53.

Shapiro, L.A. 1997. Junk representations. *British Journal for the Philosophy of Science* .

2.2e Externalism and Computation -----

Andler, D. 1995. Can we knock off the shackles of syntax? In (E. Villanueva, ed) *_Content_*. Ridgeview.

Butler, K. 1998. Content, computation, and individuation. *Synthese* 114:277-92.

Egan, F. 1995. Computation and content. *Philosophical Review* 104:181-203.

Egan, F. 1999. In defence of narrow mindedness. *Mind and Language* 14:177-94.

Kazez, J.R. 1994. Computationalism and the causal role of content. *Philosophical Studies* 75:231-60.

Kobes, B. 1990. Individualism and artificial intelligence. *Philosophical Perspectives* 4:429-56.
Winograd's SHRDLU doesn't support individualism: its concepts are anchored (to a fictional world) via its programmer, and it could have made errors.

Miscevic, N. 1996. Computation, content, and cause. *Philosophical Studies* 82:241-63.

Peacocke, C. 1995. Content, computation, and externalism. In (E. Villanueva, ed) *_Content_*. Ridgeview.

Peacocke, C. 1999. Computation as involving content: A response to Egan. *Mind and Language* 14:195-202.

Seager, W.E. 1992. Thought and syntax. *Philosophy of Science Association* 1992, 1:481-91.
Syntax is extrinsically determined, as well as semantics. So if broad content is irrelevant to psychology, syntax is too.

Wilson, R.A. 1994. Wide computationalism. *Mind* 103:351-72.

2.2f Externalism and Self-Knowledge

Berg, J. 1998. First-person authority, externalism, and wh-knowledge.
Dialectica 52:41-44.

Bernecker, S. 1996. Davidson on first-person authority and externalism.
Inquiry 39:121-39.

Bernecker, S. 1996. Externalism and the attitudinal component of
self-knowledge. *Nous* 30:262-75.

Bernecker, S. 1998. Self-knowledge and closure. In (P. Ludlow & N. Martin,
eds) *Externalism and Self-Knowledge*. CSLI.

Bilgrami, A. 1992. Can externalism be reconciled with self-knowledge?
Philosophical Topics 20:233-68.

Boghossian, P. 1989. Content and self-knowledge. *Philosophical Topics*
17:5-26.

We can't know our thought-contents by inference (circular), by introspection
(because they're relational), or directly, so we can't know them at all.

Boghossian, P. 1992. Externalism and inference. *Philosophical Issues* 2:11-28.

Boghossian, P. 1994. The transparency of mental content. *Philosophical
Perspectives* 8:33-50.

Boghossian, P. 1997. What the externalist can know a priori. *Proceedings of
the Aristotelian Society* 97:161-75.

Brown, J. 1995. The incompatibility of anti-individualism and privileged
access. *Analysis* 55:149-56.

Brown, J. 2000. Critical reasoning, understanding and self-knowledge.
Philosophy and Phenomenological Research 61:659-676.

Brueckner, A. 1990. Scepticism about knowledge of content. *Mind* 99:447-51.

Brueckner, A. 1992. What an anti-individualist knows a priori. *Analysis*
52:111-18.

Contra McKinsey 1991, anti-individualism doesn't lead to a priori knowledge.
The belief that water is wet doesn't conceptually entail facts about the
external world (e.g. H₂O), although it may metaphysically necessitate them.

Brueckner, A. 1992. Semantic answers to skepticism. *Pacific Philosophical
Quarterly* 73:200-19.

Brueckner, A. 1993. Skepticism and externalism. *Philosophia* 22:169-71.

Brueckner, A. 1994. Knowledge of content and knowledge of the world.
Philosophical Review:103-327-43.

Brueckner, A. 1995. Trying to get outside your own skin. *Philosophical Topics*
23:79-111.

Brueckner, A. 1997. Externalism and memory. *Pacific Philosophical Quarterly*
78:1-12.

Brueckner, A. 1997. Is scepticism about self-knowledge incoherent? *Analysis*
4:287-90.

Brueckner, A. 2000. Externalism and the a prioricity of self-knowledge.
Analysis 60:132-136.

Burge, T. 1988. Individualism and self-knowledge. *Journal of Philosophy* 85:649-63.

Knowledge of our thoughts is compatible with externalism: its content is self-referential and self-verifying. We needn't be able to explicate the content or its enabling conditions, or rule out twin possibilities.

Burge, T. 1996. Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society* 96:91-116.

Burge, T. 1998. Memory and self-knowledge. In (P. Ludlow & N. Martin, eds) *_Externalism and Self-Knowledge_*. CSLI.

Butler, K. 1997. Externalism, internalism, and knowledge of content. *philosophy and Phenomenological Research* 57:773-800.

Butler, K. 1998. Externalism and skepticism. *Dialogue* 37:13-34.

Chase, J. 2001. Is externalism about content inconsistent with internalism about justification? *Australasian Jouenal of Philosophy* 79:227-46.

Davidson, D. 1987. Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*.

Davies, M. 1998. Externalism, architecturalism, and epistemic warrant. In (C. Wright, B. Smith, and C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.

Ebbs, G. 1996. Can we take our words at face value? *Philosophy and Phenomenological Research* 56:499-530.

Edwards, J. 1998. The simple theory of colour and the transparency of sense experience. In (C. Wright, B. Smith, and C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.

Falvey, K. & Owens, J. 1994. Externalism, self-knowledge, and skepticism. *Philosophical Review* 103:107-37.

Falvey, K. 2000. The compatibility of anti-individualism and privileged access. *Analysis* 60:137-142.

Gallois, A. 1994. Deflationary self-knowledge. In (M. Michael & J. O'Leary-Hawthorne, eds) *_Philosophy in Mind: The Place of Philosophy in the Study of Mind_*. Kluwer.

Gallois, A. & O'Leary-Hawthorne, J. 1996. Externalism and skepticism. *Philosophical Studies* 81:1-26.

Externalist anti-skeptical arguments fail as they require us to know a priori that our terms designate natural kinds, and also because they require us to know a priori that externalism is true. A thorough analysis.

Georgalis, N. 1990. No access for the externalist: Discussion of Heil's "Privileged access". *Mind* 100:101-8.

Georgalis, N. 1994. Asymmetry of access to intentional states. *Erkenntnis* 40:185-211.

Gibbons, J. 1996. Externalism and knowledge of content. *Philosophical Review* 105:287-310.

Gibbons, J. 2001. Externalism and knowledge of the attitudes. *Philosophical Quarterly* 51:13-28.

Glock, H.J. & Preston, J.M. 1995. Externalism and first-person authority. *Monist* 78:515-33.

- Goldberg, S. 1997. Self-ascription, self-knowledge, and the memory argument. *Analysis* 57:211-19.
- Goldberg, S. 1999. The relevance of discriminatory knowledge of content. *Pacific Philosophical Quarterly* 80:136-56.
- Goldberg, S. 1999. The psychology and epistemology of self-knowledge. *Synthese* 118:165-201.
- Goldberg, S. 2000. Externalism and authoritative knowledge of content: A new incompatibilist strategy. *Philosophical Studies* 100:51-79.
- Hall, L. 1998. The self-knowledge that externalists leave out. *Southwest Philosophy Review* 14.
- Heal, J. 1998. Externalism and memory. *Aristotelian Society Supplement* 72:77-94.
- Heil, J. 1988. Privileged access. *Mind* 98:238-51.
- Kobes, B. 1996. Mental content and hot self-knowledge. *Philosophical Topics* 24:71-99.
- LePore, E. 1990. Subjectivism and environmentalism. *Inquiry* 33:197-214.
Subjectivism and environmentalism seem to clash on knowledge of content, but it's OK: under environmentalism we still know our contents w/o evidence.
- Ludlow, P. 1995. Externalism, self-knowledge, and the prevalence of slow-switching. *Analysis* 55:45-49.
Argues that cases of switching between language communities are quite common, so that Warfield's case for externalist self-knowledge doesn't work.
- Ludlow, P. 1995. Social externalism, self-knowledge, and memory. *Analysis* 55:157-59.
- Ludlow, P. 1995. Social externalism and memory: A problem? *Acta Analytica* 10:69-76.
- Ludlow, P. 1997. On the relevance of slow switching. *Analysis* 57:285-86.
- Ludlow, P. & Martin, N. 1998. *_Externalism and Self-Knowledge_*. CSLI.
- Macdonald, C. 1995. Externalism and first-person authority. *Synthese* 104:99-122.
On reconciling externalism with the non-evidential character of first-person knowledge.
- Macdonald, C. 1998. Externalism and authoritative self-knowledge. In (C. Wright, P. Smith, & C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.
- Macdonald, C., Smith, P. & Wright, C. 1998. *_Knowing Our Own Minds: Essays in Self-Knowledge_*. Oxford University Press.
- McKinsey, M. 1987. Apriorism in the philosophy of language. *Philosophical Studies* 52:1-32.
Argues that we can know the meaning of our words a priori. Analyzes twin earth cases by separating propositional meaning from linguistic meaning, which is indexical, fixes reference, and is knowable a priori.
- McKinsey, M. 1991. Anti-individualism and privileged access. *Analysis* 51:9-16.
Contra Burge: if there are conceptual connections between wide contents and

and the external world, then we can't know wide contents a priori, as otherwise we could know a priori that the world exists.

McKinsey, M. 1994. Accepting the consequences of anti-individualism. *Analysis* 54:124-8.

Reply to Brueckner 1992: The claim that belief metaphysically necessitate external facts is trivial. Almost all states do that, for Kripkean reason.

McLaughlin, B.P. & Tye, M. 1998. Externalism, Twin Earth, and self-knowledge. In (C. Macdonald, P. Smith, & C. Wright, eds) *_Knowing Our Own Minds: Essays in Self-Knowledge_*. Oxford University Press.

McLaughlin, B.P., & Tye, M. 1998. Is content-externalism compatible with privileged access? *Philosophical Review* 107:349-380.

Miller, R.W. 1997. Externalist self-knowledge and the scope of the a priori. *Analysis* 57:67-74.

Peacocke, C. 1996. Entitlement, self-knowledge, and conceptual redeployment. *Proceedings of the Aristotelian Society* 96:117-58.

Raffman, D. 1998. First-person authority and the internal reality of beliefs. In (C. Wright, B. Smith, & C. Macdonald, eds.) *_Knowing Our Own Minds_*. Oxford University Press.

Sawyer, S. 1998. Privileged access to the world. *Australasian Journal of Philosophy* 76:523-533.

Schiffer, S. 1992. Boghossian on externalism and inference. *Philosophical Issues* 2:29-38.

Szubka, T. 2000. Meaning rationalism, a priori, and transparency of content. *Philosophical Psychology* 13:491-503.

Tye, M. 1998. Externalism and memory. *Aristotelian Society Supplement* 72:77-94.

Warfield, T.A. 1992. Privileged self-knowledge and externalism are compatible. *Analysis* 52:232-37.

Boghossian's argument that externalism threatens self-knowledge fails: twin cases needn't be relevant alternatives (unless they are actual), so they don't threaten knowledge of content, by the usual standards of knowledge.

Warfield, T.A. 1995. Knowing the world and knowing our minds. *Philosophy and Phenomenological Research*.

Argues that externalism and self-knowledge imply the falsity of skepticism (though externalism alone does not). And arguments against externalist self-knowledge are no better than standard skeptical arguments.

Warfield, T.A. 1997. Externalism, privileged self-knowledge, and the irrelevance of slow switching. *Analysis* 57:282-84.

Wyler, T. 1994. First-person authority and singular thoughts. *Zeitschrift fur Philosophie Forschung* 48:585-94.

2.2g The Status of Narrow Content

Adams, F., Drebusenko, D., Fuller, G. & Stecker, R. 1990. Narrow content: Fodor's folly. *Mind and Language* 5:213-29.

Traces and criticizes Fodor's position on narrow content. Argues that narrow content isn't content, and doesn't explain behavior. Fun but arguable.

Adams, F. & Fuller, G. 1992. Names, contents, and causes. *Mind and Language* 7:205-21.

Argues that problems with names don't require an appeal to narrow content in explanation. Broad content plus associated descriptions will do the job.

Antony, L. 1989. Semantic anorexia: On the notion of content in cognitive science. In (G. Boolos, ed) *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.

Representational cognitive science has no need for narrow content -- wide contents and formal properties can do all the work. Argues that the semantics of mental expressions needn't mirror the semantics of language.

Aydede, M. 1997. Has Fodor really changed his mind on narrow content? *Mind and Language* 12:422-58.

Baker, L.R. 1985. A farewell to functionalism. *Philosophical Studies* 48:1-14. Argues that type-identical functional states can differ in narrow content, so methodological solipsism fails. Uses the example of identical programs for playing chess and arms negotiations.

Baker, L.R. 1985. Just what do we have in mind? *Midwest Studies in Philosophy* 10:25-48.

Some implausible twin cases trying to show that mental life can vary wildly while preserving physical/computational state. Bizarre.

Baker, L.R. 1986. Content by courtesy. *Journal of Philosophy* 84:197-213.

Baker, L.R. 1987. *Saving Belief*. Princeton University Press.

Lots of arguments against narrow content. Very stimulating, though wrong.

Biro, J.I. 1992. In defense of social content. *Philosophical Studies* 67:277-93.

Contra Loar 1988, the contents of "that"-clauses often reflects psychological content, even if it sometimes does not. We don't need narrow content.

Block, N. 1991. What narrow content is not. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

There are big problems specifying the "mapping" and the relevant contexts for Fodor's theory noncircularly. Narrow content either collapses into syntax or is too coarse-grained. Nontrivial narrow content must be holistic.

Block, N. 1995. Ruritania revisited. In (E. Villanueva, ed) *Content*. Ridgeview.

Brown, C. 1993. Belief states and narrow content. *Mind and Language* 8:343-67. Criticizes the "bracketing" strategy of Stich and Walker, and argues that intrinsic belief state should be individuated according to how it embeds in different environments. With a comparison with Fodor's related theory.

Chalmers, D.J. 1994. The components of content. Manuscript.

Argues for a two-dimensional intensional theory, with different kinds of intensions constituting notional and relational content. Notional content governs the dynamics of thought and behavior, and is primary in explanation.

Davies, M. 1986. Externality, psychological explanation, and narrow content. *Proceedings of the Aristotelian Society* 60:263-83.

Comments on Fodor 1987. Fodor doesn't make a conclusive case against externalism; but narrow content may be promising, and inexpressibility doesn't pose any real problems. With comparisons to neo-Fregean theories.

Dennett, D.C. 1983. Beyond belief. In (A. Woodfield, ed) *Thought and Object*. Oxford University Press. Reprinted in *The Intentional Stance* (MIT Press, 1987).

What matters are not propositional attitudes but notional attitudes; but it's hard to calibrate notional worlds. Very nice.

Devitt, M. 1990. The narrow representational theory of mind. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Not syntactic psychology nor wide psychology, but narrow psychology.

Field, H. 1989. "Narrow" aspects of intentionality and the information-theoretic approach to content. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Fodor, J.A. 1987. Individualism and supervenience. In *Psychosemantics*. MIT Press.

Science taxonomizes by causal powers, which are locally supervenient, so psychology needs a narrow notion of content. Proposes that a relativized notion -- a function from context to extension -- can do the job. Nice.

Jackson, F., and Pettit, P. 1993. Some content is narrow. In (J. Heil and A. Mele, eds) *Mental Causation*. Oxford University Press.

Argues that folk psychology needs a notion of narrow content to provide robust predictive behavioral generalizations that covers doppelgangers. If not, then some behavioral patterns would be flukey.

LePore, E. & Loewer, B. 1986. Solipsistic semantics. *Midwest Studies in Philosophy* 10:595-614.

There's no good way to construe narrow content. Phenomenologist strategy is intrinsically wide, and indexicalist strategy can't specify content.

LePore, E. & Loewer, B. 1989. Dual aspect semantics. In (S. Silvers, ed) *ReRepresentation*. Kluwer.

Loar, B. 1987. Social content and psychological content. In (R. Grimm & D. Merrill, eds) *Contents of Thought*. University of Arizona Press.

Uses examples to argue that psychological content is not fixed by the content of "that"-clauses in belief ascription, and vice versa. We require a subtler kind of narrow content to capture what's going on.

Loar, B. 1987. Subjective intentionality. *Philosophical Topics* 15:89-124.

Maloney, J.C. 1991. Saving psychological solipsism. *Philosophical Studies* 61:267-83.

Contests the "provoked/aggravated assault" example of Baker 1986. If they're doppelgangers, then their narrow content can't differ.

Manfredi, P. 1993. Two routes to narrow content: both dead ends. *Philosophical Psychology* 6:3-22.

McDermott, M. 1986. Narrow content. *Australasian Journal of Philosophy* 64:277-88.

Narrow beliefs are de re beliefs about our inputs and outputs.

McGilvray, J. 1998. Meanings are syntactically individuated and found in the head. *Mind and Language* 13:225-280.

Putnam, H. 1987. Fodor and Block on narrow content. In *Representation and Reality*. MIT Press.

Against perceptual-prototype and conceptual-role accounts of narrow content.

Quillen, K. 1986. Propositional attitudes and psychological explanation. *Mind and Language* 1:133-57.

Can't get a 'mode of presentation' account of narrow content to work, either through description theory or prototypes. Psych should be non-individualist.

- Recanati, F. 1990. Externalism and narrow content. *Nous*.
There are levels of narrowness, varying by whether independence is of actual or normal environment. Argues that this can be consistent with externalism.
- Recanati, F. 1994. How narrow is narrow content? *Dialectica* 48:209-29.
- Schiffer, S. 1989. Fodor's character. In (E. Villanueva, ed) *_Information, Semantics, and Epistemology_*. Blackwell.
- Segal, G. 2000. *_A Slim Book about Narrow Content_*. MIT Press.
- Silverberg, A. 1995. Narrow content: A defense. *Southern Journal of Philosophy* 33:109-27.
- Stalnaker, R.C. 1990. Narrow content. In (C.A. Anderson & J. Owens, eds) *_Propositional Attitudes_*. CSLI.
On some problems with narrow content, contra Loar 1987. Narrow content is hard to spell out with "diagonal" propositions. Loar doesn't show that psychological content is narrow. With some remarks on privileged access.
- Stich, S.P. 1991. Narrow content meets fat syntax. In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell.
Argues that narrow content is still too coarse-grained for explanation, classifying psychologically distinct states together. Use syntax instead.
- Taylor, K. 1989. Supervenience and levels of meaning. *Southern Journal of Philosophy* 27:443-58.
Argues that the partial character construal of narrow content is not interestingly semantic. It collapses into syntax or phenomenology.
- Taylor, K. 1989. Narrow content functionalism and the mind-body problem. *Nous* 23:355-72.
Uses a "fraternal twin earth" thought experiment to show that even de dicto attributions don't supervene on narrow role, and that narrow content can't be explicated descriptively unless it collapses into phenomenalism.
- Vaughan, R. 1989. Searle's narrow content. *Ratio* 2:185-90.
- White, S. 1982. Partial character and the language of thought. *Pacific Philosophical Quarterly* 63:347-65.
Replies to Burge/Stich arguments by introducing partial character -- a function from context to content, analogous to Kaplan's character -- as the semantic property determined by functional state and relevant to explanation.
- White, S. 1992. Narrow content and narrow interpretation. In *_The Unity of the Self_*. MIT Press.
Argues for an account of narrow content in terms of notional worlds, by considering "objective optimality" across worlds. This allows for a sort of narrow radical interpretation. With arguments against Stalnaker.
- Williams, M. 1990. Social norms and narrow content. *Midwest Studies in Philosophy* 15:425-462.
Narrow content theories can't handle the normativity of content. In-depth treatment of Burge cases and of the failures of causal and conceptual-role accounts. Normativity is fundamentally social. A long, interesting paper.
- Williamson, T. 1998. The broadness of the mental: Some logical issues. *Philosophical Perspectives* 12:389-410.

2.2h Miscellaneous -----

- Brook, D. 1992. Substantial mind. *South African Journal of Philosophy*

1:15-21.

Brown, D.J. 1993. Swampman of La Mancha. Canadian Journal of Philosophy 23:327-48.

An entertaining fable about a swampthing doppelganger of a murder witness. Does he have content? With plot twists about personal identity.

Brown, D.J. 1996. A furry tile about mental representation. Philosophical Quarterly 185:448-66.

Buekens, F. 1994. Externalism, content, and causal histories. Dialectica 48:267-86.

de Vries, W.A. 1996. Experience and the swamp creature. Philosophical Studies 82:55-80.

Argues that a swampthing isn't intelligent or intentional, with different physiological processes and no sensations, as these are functional kinds.

Edwards, S. 1994. _Externalism in the Philosophy of Mind_. Avebury.

Engel, P. 1987. Functionalism, belief, and content. In (Torrance, ed) _The Mind and the Machine_. Horwood.

Gauker, C. 1991. Mental content and the division of epistemic labour. Australasian Journal of Philosophy 69:302-18.

Gibbons, J. 1993. Identity without supervenience. Philosophical Studies 70:59-79.

Houghton, D. 1997. Mental content and external representations: internalism, anti-internalism. Philosophical Quarterly 47:159-77.

Jackson, F. & Pettit, P. 1988. Functionalism and broad content. Mind 97:318-400.

Should construe functionalism broadly rather than narrowly; then can handle the problem of broad content.

Katz, J. 1990. The domino theory. Philosophical Studies 58:3-39.

Anti-intensional arguments are not independent but a series of dominos. Quine/Quine/Davidson/Putnam/Burge rise and fall together.

Macdonald, C. 1990. Weak externalism and mind-body identity. Mind 99:387-404.

McCulloch, G. 1995. _The Mind and its World_. Routledge.

McGinn, C. 1982. The structure of content. In (A. Woodfield, ed) _Thought and Object_. Oxford University Press.

Belief content has two distinct elements, one causal-explanatory, the other truth-related.

Owens, J. 1987. In defense of a different Doppelganger. Philosophical Review 96:521-54.

Owens, J. 1992. Psychophysical supervenience: Its epistemological foundation. Synthese 90:89-117.

Pereboom, D. 1995. Conceptual structure and the individuation of content. Philosophical Perspectives 9:401-428.

Preti, C. 2000. Belief and desire under the elms. Protosociology 14:270-284.

Rey, G. 1992. Semantic externalism and conceptual competence. Proceedings of the Aristotelian Society 66:315-33.

Supplements externalist "locking" theories of content with an account of

internal "conceptions" by which thoughts lock onto environmental kinds, with that aid of dthat operators, thus solving various philosophical problems.

Rowlands, M. 1995. Externalism and token-token identity. *Philosophia* 24:359-75.

Rowlands, M. 1999. *_The Body in Mind: Understanding Cognitive Processes_*. Cambridge University Press.

Rudd, A. 1997. Two types of externalism. *Philosophical Quarterly* 47:501-7.

Seager, W.E. 1992. Externalism and token identity. *Philosophical Quarterly* 42:439-48.

Stalnaker, R.C. 1989. On what's in the head. *Philosophical Perspectives* 3:287-319.

Thomas, J. 1996. Analogies and the mind of the replica: Sunburn, the little green bug, and the fake plant. *Philosophical Quarterly* 46:364-371.

Walker, V. 1990. In defense of a different taxonomy: A reply to Owens. *Philosophical Review* 99.

Contra Owens 1987: wide intentional descriptions and molar bodily descriptions don't exhaust the options. A bracketing strategy gives a narrow intentional taxonomy of mental states.

Williams, M. 1990. Externalism and the philosophy of mind. *Philosophical Quarterly* 40:352-80.

Woodfield, A. 1986. Two categories of content. *Mind and Language* 1:319-54.

2.3 Causal Theories of Content

2.3a Information-Based Accounts (Dretske, etc)

Barwise, J. & Perry, J. 1983. *_Situations and Attitudes_*. MIT Press.

Barwise, J. 1986. Information and circumstance. *Notre Dame Journal of Formal Logic*.

Defending information against Fodor 1986. Information is objective but relational, and depends on the relevant constraints between representation and environment. Circumstances play a vital role.

Barwise, J. 1987. Unburdening the language of thought. *Mind and Language*.

Bogdan, R.J. 1988. Information and semantic cognition: An ontological account. *Mind and Language*.

From material (formal) info to semantic info via teleology; from semantic information to representation via internal structure. Cute. With a good reply by Israel, and a terse reply by Dretske.

Bogdan, R.J. 1987. Mind, content and information. *Synthese*.

Clark, A. 1993. Mice, shrews, and misrepresentation. *Journal of Philosophy* 90:290-310.

Uses information theory to analyze misrepresentation. A signal represents what it carries most information about, not what it correlates best with. Treating some signals as noise can increase information content.

Coulter, J. 1995. The informed neuron: Issues in the use of information theory in the behavioral sciences. *Minds and Machines* 5:583-96.

- Dretske, F. 1981. *Knowledge and the Flow of Information*. MIT Press.
Defines knowledge content is in terms of information-flow from events, and applies to various aspects of psychology.
- Dretske, F. 1983. *Precis of Knowledge and the Flow of Information*. Behavioral and Brain Sciences 6:55-90.
A summary of the book, with commentary and replies.
- Dretske, F. 1990. Putting information to work. In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.
On the causal role of information (as opposed to meaning). Information is causally efficacious if considered with respect to learning. With commentary by Brian Smith.
- Fodor, J.A. 1986. Information and association. Notre Dame Journal of Formal Logic 27.
Transmission of information is no good without the encoding of information. With criticisms of associative networks, which transmit without encoding, and criticism of Barwise & Perry's account of attunement to a relation.
- Fodor, J.A. 1987. A situated grandmother. Mind and Language.
- Foley, R. 1987. Dretske's 'information-theoretic' account of knowledge. Synthese.
- Gjelsvik, O. 1991. Dretske on knowledge and content. Synthese 86:425-41.
- Grandy, R. 1987. Information-based epistemology, ecological epistemology and epistemology naturalized. Synthese 70:191-203.
Shannon's notion of information is more useful for naturalized epistemology than Dretske's.
- Hardcastle, V.G. 1994. Indicator semantics and Dretske's function. Philosophical Psychology 7:367-82.
- Heller, M. 1991. Indication and what might have been. Analysis 51:187-91.
We need to analyze indication in terms of "close enough" worlds; the relevant conditionals are "might"-conditionals.
- Israel, D. & Perry, J. 1990. What is information? In (P. Hanson, ed) *Information, Language and Cognition*. University of British Columbia Press.
- Jackendoff, R. 1985. Information is in the mind of the beholder. Linguistics and Philosophy 8:23-33.
Argues that a representationalist theory of semantics beats a realist one.
- Loewer, B. 1987. From information to intentionality. Synthese.
- Morris, W.E. 1990. The regularity theory of information. Synthese 82:375-398.
Dretske has problems with ruling out alternative possibilities; and there is a gap between information-caused belief and knowledge.
- Savitt, S. 1987. Absolute informational content. Synthese 70:185-90.
Makes a distinction between absolute information and information that's relative to other knowledge.
- Sayre, K.M. 1986. Intentionality and information processing: An alternative model for cognitive science. Behavioral and Brain Sciences 9:121-38.
- Sayre, K.M. 1987. Cognitive science and the problem of semantic content. Synthese 70:247-69.
On problems with a computational approach to content: computers process info(t), the non-semantic content of communication theory, not info(s), or

semantic content. Get info(s) from efficient processing of mutual info(t).

Sturdee, D. 1997. The semantic shuffle: Shifting emphasis in Dretske's account of representational content. *Erkenntnis* 47:89-104.

Taylor, K. 1987. Belief, information and semantic content: A naturalist's lament. *Synthese* 71:97-124.

Usher, M. 2001. A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language* 16:331-334.

Winograd, T. 1987. Cognition, attunement and modularity. *Mind and Language*.

Zalabardo, J.L. 1995. A problem for information-theoretic semantics. *Synthese* 105:1-29.

2.3b Asymmetric Dependence (Fodor)

Fodor, J.A. 1987. Meaning and the world order. In *_Psychosemantics_*. MIT Press.

Defends and refines a causal theory, using the notion of asymmetric dependence of a token upon the world.

Fodor, J.A. 1990. A theory of content II. In *_A Theory of Content_*. MIT Press.

Defending the asymmetric dependence theory against various objections.

Adams, F. & Aizawa, K. 1992. 'X' means X: Semantics Fodor-style. *Minds and Machines* 2:175-83.

Adams, F. & Aizawa, K. 1993. Fodorian semantics, pathologies, and "Block's problem". *Minds and Machines* 3:97-104.

Adams, F. & Aizawa, K. 1994. 'X' means X: Fodor/Warfield semantics. *Minds and Machines* 4:215-31.

Adams, F. & Aizawa, K. 1997. Fodor's asymmetric causal dependency theory and proximal projections. *Southern Journal of Philosophy* 35:433-437.

Antony, L. & Levine, J. 1991. The nomic and the robust. In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell.

Baker, L.R. 1990. On a causal theory of content. *Philosophical Perspectives*.

Baker, L.R. 1991. Has content been naturalized? In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell.

Bernier, P. 1993. Narrow content, context of thought, and asymmetric dependence. *Mind and Language* 8:327-42.

Boghossian, P. 1991. Naturalizing content. In (B. Loewer & G. Rey, eds) *_Meaning in Mind: Fodor and his Critics_*. Blackwell.

Argues that Fodor's theory is a type-1 theory, requiring naturalistically specifiable circumstances in which a symbol is only caused by its referent; and that these theories fail for various reasons, e.g. verificationism.

Cram, H-R. 1992. Fodor's causal theory of representation. *Philosophical Quarterly* 42:56-70.

Fodor's theory has counterexamples and can't explain its counterfactuals; but we can save it by borrowing from Dretske's account of misrepresentation.

Gibson, M. 1996. Asymmetric dependencies, ideal conditions, and meaning. *Philosophical Psychology* 9:235-59.

Loar, B. 1991. Can we explain intentionality? In (B. Loewer & G. Rey, eds) Meaning in Mind: Fodor and his Critics. Blackwell.

Maloney, J.C. 1990. Mental misrepresentation. *Philosophy of Science* 57:445-58.

Manfredi, P.A. & Summerfield, D.M. 1992. Robustness without asymmetry: A flaw in Fodor's theory of content. *Philosophical Studies* 66:261-83.

Rupert, R. 2000. Dispositions indisposed: Semantic atomism and Fodor's theory of content. *Pacific Philosophical Quarterly* 81:325-349.

Seager, W.E. 1993. Fodor's theory of content: problems and objections. *Philosophy of Science* 60:262-77.

Wallis, C. 1995. Asymmetric dependence, representation, and cognitive science. *Southern Journal of Philosophy* 33:373-401.

Warfield, T.A. 1994. Fodorian semantics: A reply to Adams and Aizawa. *Minds and Machines* 4:205-14.

2.3c Causal Accounts, General -----

Aizawa, K. 1994. Lloyd's dialectical theory of representation. *Mind and Language* 9:1-24.

Cummins, R. 1989. Representation and covariation. In (S. Silvers, ed) ReRepresentation. Kluwer.

Cummins, R. 1997. The LOT of the causal theory of mental content. *Journal of Philosophy* 94:535-542.

Fodor, J.A. 1984. Semantics, Wisconsin style. *Synthese* 59:231-50. Reprinted in RePresentations (MIT Press, 1980).

A somewhat sympathetic commentary on the Dretske/Stampe causal theories, but raising the problem of misrepresentation.

Fodor, J.A. 1990. Information and representation. In (P. Hanson, ed) Information, Language and Cognition. University of British Columbia Press.

Godfrey-Smith, P. 1989. Misinformation. *Canadian Journal of Philosophy* 19:533-50.

On various attempts to solve the error problem and why they don't work.

Godfrey-Smith, P. 1991. Signal, decision, action. *Journal of Philosophy* 88:709-22.

World-head reliability is just as important as head-world reliability. With arguments and examples from signal detection theory.

Jacquette, D. 1996. Lloyd on intrinsic natural representation in simple mechanical minds. *Minds and Machines* 6:47-60.

Maloney, J.C. 1994. Content: Covariation, control, and contingency. *Synthese* 100:241-90.

McLaughlin, B.P. 1987. What is wrong with correlational psychosemantics. *Synthese*.

Ray, G. 1997. Fodor and the inscrutability problem. *Mind and Language* 12:475-89.

Stampe, D. 1977. Towards a causal theory of linguistic representation.

Midwest Studies in Philosophy 2:42-63.

Stampe, D. 1986. Verificationism and a causal account of meaning. *Synthese* 69:107-37.

Stampe, D. 1991. Content, context, and explanation. In (E. Villanueva, ed) *_Information, Semantics, and Epistemology_*. Blackwell.

Viger, C.D. 2001. Locking on to the language of thought. *Philosophical Psychology* 14:203-215.

Warmbrod, K. 1992. Primitive representation and misrepresentation. *Topoi* 11:89-101.

Weitzman, L. 1996. What makes a causal theory of content anti-skeptical? *Philosophy and Phenomenological Research* 56:299-318.

2.3d Teleological Approaches (Millikan, etc)

Adams, F. & Aizawa, K. 1997. Rock beats scissors: Historicalism fights back. *Analysis* 57:273-81.

Agar, N. 1993. What do frogs really believe? *Australasian Journal of Philosophy* 71:1-12.

Argues that a teleological account can resolve content indeterminacies, by an appeal to counterfactuals in examining what properties were selected for.

Bogdan, R. 1994. *_Grounds for Cognition: How Goal-Guided Behavior Shapes the Mind_*. Lawrence Erlbaum.

Braddon-Mitchell, D. & Jackson, F. 1997. The teleological theory of content. *Australasian Journal of Philosophy* 75:474-89.

Clarke, M. 1996. Darwinian algorithms and indexical representation. *Philosophy of Science* 63:27-48.

Dennett, D.C. 1988. Fear of Darwin's optimizing rationale. Manuscript.
Defends evolutionary theories of content against Fodor.

Dennett, D.C. 1988. Evolution, error and intentionality. In *_The Intentional Stance_*. MIT Press.

Attacks original intentionality (Fodor/Burge/Dretske/Searle/Kripke) -- our intentionality, if anything, is derived through evolution, and so is as indeterminate as that of an artifact.

Dretske, F. 1986. Misrepresentation. In (R. Bogdan, ed) *_Belief: Form, Content, and Function_*. Oxford University Press.

Tries to deal with misrepresentation by appealing to function.

Elder, C.L. 1998. What versus how in naturally selected representations. *Mind* 107:349-363.

Fodor, J.A. 1990. Psychosemantics, or, Where do truth conditions come from? In (W. Lycan, ed) *_Mind and Cognition_*. Blackwell.

Truth conditions are "entry conditions" for belief under "normal function". Later repudiated.

Fodor, J.A. 1990. A theory of content I. In *_A Theory of Content_*. MIT Press.

Teleological solutions can't work, because of underdetermination and so on.

Keeley, B. 1999. Fixing content and function in neurobiological systems: The neuroethology of electroreception. *Biology and Philosophy* 14:395-430.

Lalor, B.J. 1998. Swampman, etiology, and content. *Southern Journal of Philosophy* 36:215-232.

Levine, J. 1996. Swampjoe: mind or simulation? *Mind and Language* 11:86-91.

Macdonald, G. 1989. Biology and representation. *Mind and Language* 4:186-200.

Matthen, M. 1988. Biological functions and perceptual content. *Journal of Philosophy* 85:5-27.

Millikan, R.G. 1979. An evolutionist approach to language. *Philosophy Research Archives* 5.

Millikan, R.G. 1984. *Language, Thought and Other Biological Categories*. MIT Press.

An evolutionary account of thought, content, and various intentional phenomena, appealing to proper functions and adaptational role to individuate contents.

Millikan, R.G. 1986. Thoughts without laws: Cognitive science with content. *Philosophical Review* 95:47-80.

The content of a desire is its adaptational Proper Function; the content of a belief is its Normal Condition for success.

Millikan, R.G. 1989. Biosemantics. *Journal of Philosophy* 86:281-97.

Representation content is determined by the consumption of a representation, not its production. The representation-world correspondence is best taken as a normal condition for the consumer's function.

Millikan, R.G. 1989. In defense of proper functions. *Philosophy of Science* 56:288-302.

Millikan, R.G. 1990. Compare and contrast Dretske, Fodor, and Millikan on teleosemantics. *Philosophical Topics* 18:151-61.

Contrasting positions on the role of representation production and consumption; also on the role of reliability, articulateness, and learning.

Millikan, R.G. 1991. Speaking up for Darwin. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

A reply to some of Fodor's criticisms of teleological theories in *Psychosemantics* and elsewhere. With some remarks on Fodor's asymmetric dependence theory.

Millikan, R.G. 1993. *White Queen Psychology and Other Essays for Alice*. MIT Press,

A collection of papers on teleological semantics and other issues about psychology and mental content.

Millikan, R.G. 1996. On swampkinds. *Mind and Language* 11:103-17.

Millikan, R.G. 1997. Troubles with Wagner's reading of Millikan. *Philosophical Studies* 86:93-96.

Neander, K. 1995. Misrepresenting and malfunctioning. *Philosophical Studies* 79:109-41.

Neander, K. 1995. Dretske's innate modesty. *Australasian Journal of Philosophy* 74:258-74.

Neander, K. 1996. Swampman meets swampcow. *Mind and Language* 11:118-29.

It's not unreasonable to deny a swampthing beliefs: swampcows aren't cows and swamphearts aren't hearts. Semantic norms are plausibly grounded in

biological norms and so in history.

Newton, N. 1992. Dennett on intrinsic intentionality. *Analysis* 52:18-23.
Contra Dennett 1988, designed creatures can have intrinsic (if not original) intentionality. Overall purpose is dependent on designer's goals, but specific contents need not be.

Papineau, D. 1984. Representation and explanation. *Philosophy of Science* 51:550-72.

A teleological theory of belief/desire contents: the satisfaction conditions for a desire are those effects for which it was selected; truth conditions for a belief are circumstances resulting in satisfaction of desires.

Papineau, D. 1990. Truth and teleology. In (D. Knowles, ed) *Explanation and its Limits*. Cambridge University Press.

Best theory is combination of a success-guaranteeing account of truth-conditions with a teleological account of desire.

Papineau, D. 1991. Teleology and mental states. *Proceedings of the Aristotelian Society* 65.

Papineau, D. 1996. Doubtful intuitions. *Mind and Language* 11:130-32.

Papineau, D. 1998. Teleosemantics and indeterminacy. *Australasian Journal of Philosophy* 76:1-14.

Papineau, D. 2001. The status of teleosemantics, or how to stop worrying about Swampman. *Australasian Journal of Philosophy* 79:279-89.

Pickles, D. 1989. Intentionality, representation, and function. *Sussex University, Cognitive Science Research Paper* 140.

Combining the analysis-relative and historical accounts of function, and using these to give an account of intentionality: representation are produced by conditional productive functions. Argues against Fodor on indeterminacy.

Pietrowski, P.M. 1992. Intentionality and teleological error. *Pacific Philosophical Quarterly* 73:267-82.

Millikan's theory has an implausible consequence: creatures' belief contents can involve properties which they cannot discriminate. With examples.

Ross, D. & Zawidzki, T. 1994. Information and teleosemantics. *Southern Journal of Philosophy* 32:393-419.

Rountree, J. 1997. The plausibility of teleological content ascriptions: A reply to Pietroski. *Pacific Philosophical Quarterly* 78:404-??.

Rowlands, M. 1996. Teleological semantics. *Mind* 106:279-304.

Rupert, R.D. 1999. Mental representations and Millikan's theory of intentional content: Does biology chase causality? *Southern Journal of Philosophy* 37:113-140.

Sehon, S.R. 1994. Teleology and the nature of mental states. *American Philosophical Quarterly* 31:63-72.

Shapiro, L. 1996. Representation from bottom to top. *Canadian Journal of Philosophy* 26:523-42.

Shapiro, L. 1992. Darwin and disjunction: Foraging theory and univocal assignments of content. *Philosophy of Science Association* 1992, 1:469-80.

Sullivan, S.R. 1993. From natural function to indeterminate content. *Philosophical Studies* 69:129-37.

Wagner, S. 1996. Teleosemantics and the troubles of naturalism. *Philosophical Studies* 82:81-110.

Teleosemantics has big problems with indeterminacy, holism, false belief, and "psychophysical normalcy" in causation. So do all naturalistic stories.

Zawidzki, T. & Ross, D. 1994. Information and teleosemantics. *Southern Journal of Philosophy* 32:393-419.

2.3e Conceptual Role Approaches

Block, N. 1986. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* 10:615-78.

An in-depth program for conceptual-role semantics, and its role in a two-factor account of meaning. Also a defense of narrow content.

Block, N. 1988. Functional role and truth conditions. *Proceedings of the Aristotelian Society* 61:157-181.

A defense of functional role semantics, and an account of its relation to truth-conditional factors. A two-factor theory will handle wide content.

Boghossian, P.A. 1994. Inferential-role semantics and the analytic/synthetic distinction. *Philosophical Studies*.

No matter how we understand the denial of the analytic/synthetic distinction, the falsity of inferential-role semantics does not follow. The meaning-constitutive inferences needn't be the analytic inferences.

Brandom, R. 1994. *Making It Explicit*. Harvard University Press.

Brandom, R. 1994. Reasoning and representing. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind*. Kluwer.

Cummins, R. 1992. Conceptual role semantics and the explanatory role of content. *Philosophical Studies* 65:103-127.

CRS conflates representation content and attitude content (which depends on a representation's "target"), so can't handle representation content; it makes all content-based explanations vacuous; and it can't handle error properly.

Field, H. 1977. Logic, meaning, and conceptual role. *Journal of Philosophy* 74:379-409.

Explicates conceptual role in terms of conditional probability, and analyzes meaning as conceptual role plus reference. With remarks on truth, descriptions, and synonymy.

Field, H. 1978. Mental representation. *Erkenntnis* 13:9-61.

Fodor, J.A. & LePore, E. 1991. Why meaning (probably) isn't conceptual role. *Mind and Language* 6:328-43.

Conceptual role semantics isn't compatible with compositional semantics and the denial of an analytic/synthetic distinction, as full conceptual roles aren't compositional, and there's no way to specify a relevant subset.

Harman, G. 1974. Meaning and semantics. In (M. Munitz & P. Unger, eds) *Semantics and Philosophy*. New York University Press.

Harman, G. 1975. Language, thought, and communication. In (K. Gunderson, ed) *Language, Mind, and Knowledge*. University of Minnesota Press.

Harman, G. 1982. Conceptual role semantics. *Notre Dame Journal of Formal Logic* 28:242-56.

Meaning and content is determined by the role of symbols in thought (e.g. inference and perception). With remarks on indeterminacy, context-dependence,

the linguistic division of labor, qualia, speech acts, and more.

Horowitz, A. 1992. Functional role and intentionality. *Theoria* 58:197-218.

Loar, B. 1982. Conceptual role and truth conditions. *Notre Dame Journal of Formal Logic* 23:272-83.

On the relation between conceptual role and truth-conditions. Contra Harman, truth-conditions are to an extent independent of conceptual role.

Loewer B. 1982. The role of 'Conceptual role semantics'. *Notre Dame Journal of Formal Logic* 23:305-15.

Contra Harman 1982, truth-conditions are central to a semantic theory.

Perlman, M. 1997. The trouble with two-factor conceptual role theories. *Minds and Machines* 7:495-513.

Silverberg, A. 1992. Putnam on functionalism. *Philosophical Studies* 67:111-31.

Argues against Putnam 1987 that conceptual role plays an important role in determining meaning. Appeals to the induction theory of Holland et al.

Toribio, J. 1997. Twin pleas: Probing content and compositionality. *Philosophy and Phenomenological Research* 57:871-89.

Warfield, T.A. 1993. On a semantic argument against conceptual role semantics. *Analysis* 53:298-304.

Contra Fodor and Lepore, meanings can be compositional even if inferential roles are not, as long as meanings only supervene on inferential role.

2.3f Theories of Content, Misc

Bestor, T.W. 1991. Naturalizing semantics: New insights or old folly? *Inquiry* 34:285-310.

Brook, A. & Stainton, R. 1997. Fodor's new theory of content and computation. *Mind and Language* 12:459-74.

Callaway, H.G. 1995. Intentionality naturalized: Continuity, reconstruction, and instrumentalism. *Dialectica* 49:147-68.

Churchland, P.M. & Churchland, P.S. 1983. Stalking the wild epistemic engine. *Nous* 17:5-18.

On "translational" (conceptual) and "calibrational" (referential) content. Relation of content issues to computational issues.

Cummins, R. 1989. *_Meaning and Mental Representation_*. MIT Press. Critiques other views, offers interpretational semantics.

Cummins, R. 1996. *_Representations, Targets, and Attitudes_*. MIT Press.

Dennett, D.C. 1991. Ways of establishing harmony. In (B. McLaughlin, ed) *_Dretske and his Critics_*. Blackwell.

On the ways in which meanings can come to cohere with their causal roles: learning, natural selection, and design. Criticizes Dretske for undervaluing the latter two: all three are in the same boat.

Dretske, F. 1986. Aspects of cognitive representation. In (M. Brand & R. Harnish, eds) *_The Representation of Knowledge and Belief_*. University of Arizona Press.

On the reference and content of representations. Reference is determined by causation; content, i.e. representation "as", is determined by functional role, when functioning normally in natural habitat.

Shapiro, L.A. 1997. The nature of nature: Rethinking naturalistic theories of intentionality. *Philosophical Psychology* 10:309-322.

Silvers, S. 1991. On naturalizing the semantics of mental representation. *British Journal for the Philosophy of Science* 42:49-73.

Stalnaker, R. 1991. How to do semantics for the language of thought. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and His Critics*. Blackwell. On some tensions in Fodor's view of content: e.g. narrow content must be dependent on functional role, which seems to lead to holism. The role of denotational semantics as a defense is unclear.

2.4 Representation (General) [see also 4.2]

Bickhard, M. 1993. Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence* 5:285-33.

Blachowicz, J. 1997. Analog representation beyond mental imagery. *Journal of Philosophy* 94:55-84.

Chomsky, N. 1980. Rules and representations. *Behavioral and Brain Sciences* 3:1-61.

Clark, A. 1995. Moving minds: Situating content in the service of real-time success. *Philosophical Perspectives* 9:89-104.

Dalenoort, G.J. 1990. Toward a general theory of representation. *Psychological Research* 52:229-237.

Fodor, J.A. 1986. Why paramecia don't have mental representations. *Midwest Studies in Philosophy* 10:3-23.
Because paramecia can't respond to non-nomic properties of the stimulus.
Perceptual categories vs. sensory manifolds.

Freeman, W. & Skarda, C.A. 1990. Representations: who needs them? In (J. McGaugh, J. Weinberger, & G. Lynch) *Brain Organization and Memory*. Guilford Press.

Gillett, G. 1989. Representations and cognitive science. *Inquiry* 32:261-77.

Goldman, A. 1986. Constraints on representation. In (M. Brand & R. Harnish, eds) *The Representation of Knowledge and Belief*. University of Arizona Press.

Grush, R. 1997. The architecture of representation. *Philosophical Psychology* 10:5-23.

Hatfield, G. 1989. Computation, representation and content in noncognitive theories of perception. In (S. Silvers, ed) *ReRepresentation*. Kluwer.

Hogan, M. 1994. What is wrong with an atomistic account of mental representation. *Synthese* 100:307-27.

Jackendoff, R. 1991. The problem of reality. *Nous* 25:411-33.
On the philosophical (inward-out) vs. psychological (outward-in) approaches to the mind-world relation; the psychological approach is more useful in understanding representation. Internal reality is an imperfect construction.

Kukla, R. 1992. Cognitive models and representation. *British Journal for the Philosophy of Science* 43:219-32.

Lloyd, D. 1987. Mental representation from the bottom up. *Synthese* 70:23-78.

Lycan, W.G. 1989. Ideas of representation. In (Weissbord, ed) Mind, Value and Culture: Essays in Honor of E.M. Adams. Ridgeview.

Matthews, R.J. 1984. Troubles with representationalism. *Social Research* 51:1065-97.

Millikan, R.G. 1995. Pushmi-pullyu representations. *Philosophical Perspectives* 9:185-200.

Richardson, R.C. 1981. Internal representation: Prologue to a theory of intentionality. *Philosophical Topics* 12:171-212.

Shanon, B. 1991. Representations -- senses and reasons. *Philosophical Psychology* 4:355-74.

On different senses of "representation" -- external, experiential, mental locus, substrate of meaning, mediating functions, technicalpsychological.

Shanon, B. 1993. The Representational and the Presentational: An Essay on Cognition and the Study of Mind. Prentice-Hall.

Sober, E. 1976. Mental representations. *Synthese* 33:101-48.

Sterelny, K. 1995. Basic minds. *Philosophical Perspectives* 9:251-70.

van Gulick, R. 1982. Mental representation: A functionalist view. *Pacific Philosophical Quarterly* 63:3-20.

On the distinction between representation and representation-use.

Wallis, C. 1994. Representation and the imperfect ideal. *Philosophy of Science* 61:407-28.

2.5 The Explanatory Role of Content (Dretske, etc)

Adams, F. 1991. Causal contents. In (B. McLaughlin, ed) Dretske and his Critics. Blackwell.

On Dretske's account of the causal role of content. Addresses some objections: Dennett's worries about intrinsic intentionality, Fodor's about external causal powers, and some worries about syntax.

Baker, L.R. 1991. Dretske on the explanatory role of belief. *Philosophical Studies* 63:99-111.

Bogdan, R.J. 1989. Does semantics run the psyche? *Philosophy and Phenomenological Research* 49:687-700.

A critique of Fodor. Semantics per se doesn't cause. Also, Fodor's is an account of the what, not the how, of semantics. Somewhat bizarre.

Cummins, R. 1991. Mental meaning in psychological explanation. In (B. McLaughlin, ed) Dretske and his Critics. Blackwell.

Criticizes Dretske's account of the role of content, especially because of its dependence on an organism's history; also, it may not cohere with work in cognitive science. Argues for an interpretational, not a causal account.

Devitt, M. 1991. Why Fodor can't have it both ways. In (B. Loewer & G. Rey, eds) Meaning in Mind: Fodor and his Critics. Blackwell.

Dretske, F. 1987. The explanatory role of content. In (R. Grimm & D. Merrill, eds) Contents of Thought. University of Arizona Press.

Content must explain why (not how) an internal state caused a certain output. The explanation is given in terms of what a state has historically indicated. With thermostats and sea-snails as examples. Comments by Cummins, and reply.

- Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. MIT Press.
- Dretske, F. 1990. Does meaning matter? In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.
- Dretske, F. 1994. Reply to Slater and Garcia-Carpintero. *Mind and Language* 9:203-8.
- Dretske, F. 1995. Reply: Causal relevance and explanatory exclusion. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.
- Dretske, F. 1996. The explanatory role of content: Reply to Melnyk and Noordhof. *Mind and Language* 11:223-29.
- Elder, C.L. 1996. Content and the subtle extensionality of "... explains ...". *Philosophical Quarterly* 46:320-32.
- Fodor, J.A. 1986. Banish DisContent. In (J. Butterfield, ed) *Language, Mind, and Logic*. Cambridge University Press. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).
- Garcia-Carpintero, M. 1994. Dretske on the causal efficacy of meaning. *Mind and Language* 9:181-202.
- Hassrick, B. 1995. Fred Dretske on the explanatory role of semantic content. *Conference* 6:59-66.
- Horgan, T. 1991. Actions, reasons, and the explanatory role of content. In (B. McLaughlin, ed) *Dretske and his Critics*. Blackwell.
Distinguishes three problems of mental causation (extrinsic factors, exclusion of the nonphysical, anomalism). Criticizes Dretske's theory (can't handle unlearned or here-and-now reasons), offers a counterfactual account.
- Melnyk, A. 1996. The prospects for Dretske's account of the explanatory role of belief. *Mind and Language* 11:203-15.
- Noordhof, P. 1996. Accidental associations, local potency, and a dilemma for Dretske. *Mind and Language* 11:216-22.
- Perry, J. & Israel, D. 1991. Fodor and psychological explanation. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.
- Pylyshyn, Z.W. 1987. What's in a mind? *Synthese* 70:97-122.
We must individuate mental states by semantics, not just by function, as we need representation to capture generalizations about behavior; particularly due to the information-sensitivity and stimulus-independence of behavior.
- Slater, C. 1994. Discrimination without indication: Why Dretske can't lean on learning. *Mind and Language* 9:163-80.
- Wallis, C. 1994. Using representation to explain. In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

2.6 Concepts -----

- Barber, A. 1998. The pleonasticity of talk about concepts. *Philosophical Studies* 89:53-86.
- Brown, H. 1986. Sellars, concepts, and conceptual change. *Synthese* 68:275-307.

Burge, T. 1993. Concepts, definitions, and meaning. *Metaphilosophy* 24:309-25.

Churchland, P.M. 1998. Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy* 95:5-32.

Cussins, A. 1990. The connectionist construction of concepts. In (M. Boden, ed) *The Philosophy of AI*. Oxford University Press.

Connectionism builds up concepts from the nonconceptual level. From nonconceptual content (e.g. perceptual experiences) to the emergence of objectivity.

Fodor, J. & Lepore, E. 1996. The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition* 58:253-70.

Fodor, J. 1995. Concepts: A potboiler. *Cognition* 50:133-51. Also in (E. Villanueva, ed) *Content*. Ridgeview.

Fodor, J. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.

Franks, B. 1992. Realism and folk psychology in the ascription of concepts. *Philosophical Psychology* 5:369-90.

Gardenfors, P. 1997. Meanings as conceptual structures. In (M. Carrier & P. Machamer, eds) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Gauker, C. 1993. An extraterrestrial perspective on conceptual development. *Mind and Language* 8:105-30.

Grandy, R.E. 1989. Concepts, prototypes, and information. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.

Jackendoff, R. 1989. What is a concept, that a person may grasp it? *Mind and Language* 4:68-102.

Khalidi, M.A. 1995. Two concepts of concept. *Mind and Language* 10:402-22.

Levine, A. & Bickhard, MH. 1999. Concepts: Where Fodor went wrong. *Philosophical Psychology* 12:5-23.

Livingston, K.R. 1989. Concepts, categories, and epistemology. *Philosophia* 19:265-300.

Neisser, U. (ed) 1981. *Concepts and Conceptual Development*. Cambridge University Press.

Osherson, D.N. & Smith, E.E. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition* 9:35-58.

Margolis, E. 1995. The significance of the theory analogy in the psychological study of concepts. *Mind and Language* 10:45-71.

Margolis, E. 1998. How to acquire a concept. *Mind and Language* 13:347-369.

Margolis, E. 1999. What is conceptual glue? *Minds and Machines* 9:241-255.

Margolis, E. & Laurence, S. 1999. *Concepts: Core Readings*. MIT Press.

Millikan, R.G. 1994. On unclear and indistinct ideas. *Philosophical Perspectives* 8:75-100.

Millikan, R.G. 1997. A common structure for concepts of individuals, stuffs, and kinds: More mama, more milk, and more mouse. *Behavioral and Brain*

Sciences.

Peacocke, C. 1989. What are concepts? *Midwest Studies of Philosophy* 14.

Peacocke, C. 1989. Possession conditions: A focal point for theories of concepts. *Mind and Language* 4:51-56.

Peacocke, C. 1991. The metaphysics of concepts. *Mind* 100:525-46.

Peacocke, C. 1992. *_A Study of Concepts_*. MIT Press.

Peacocke, C. 1996. *Precis of _A Study of Concepts_*. *Philosophy and Phenomenological Research* 56:407-52.

A symposium on the book, with comments by Heal, Rey, Papineau.

Peacocke, C. 1996. Can a theory of concepts explain the a priori: A reply to Skorupski. *International Journal of Philosophical Studies* 4:154-60.

Peacocke, C. 1996. The relation between philosophical and psychological theories of concepts. In (P. Millican & A. Clark, eds) *_Machines and Thought_*. Oxford University Press.

Peacocke, C. 2000. Theories of concepts: A wider task. *European Journal of Philosophy* 8:298-321.

Perlman, M. 2000. *_Conceptual Flux: Mental Representation, Misrepresentation, and Concept Change_*. Kluwer.

Pitt, D. 1999. In defense of definitions. *Philosophical Psychology* 12:139-156.

Ramsey, W. 1992. Prototypes and conceptual analysis. *Topoi* 11:59-70.
On the significance of psychological work on concepts for philosophical conceptual analysis -- simple, precise analyses do not exist in general.

Rey, G. 1983. Concepts and stereotypes. *Cognition* 15:237-62.

Rips, L.J. 1995. The current status of research on concept combination. *Mind and Language* 10:72-104.

Sellars, W. 1948. Concepts as involving laws and inconceivable without them. *Philosophy of Science* 15:287-313.

Sellars, W. 1974. Conceptual change. In *_Essays in Philosophy and its History_*. Reidel.

Smith, E.E. & Medin, D.L. 1981. *_Categories and Concepts_*. Harvard University Press.

Sosa, E. 1993. Abilities, concepts, and externalism. In (J. Heil & A. Mele, eds) *_Mental Causation_*. Oxford University Press.
On concepts as abilities, and on construals of abilities that lead to internalism and externalism. Maybe the relevant abilities are characterized externally but determined internally. Remarks on Putnam, Davidson, Burge.

Thagard, P. 1990. Concepts and conceptual change. *Synthese* 82:255-74.

van Brakel, J. 1991. Meaning, prototypes, and the future of cognitive science. *Minds and Machines* 1:233-57.

Vision, G. 2001. Flash! Fodor splits the atom. *Analysis* 61:5-10.

Weitz, M. 1988. *_Theories of Concepts: A History of the Major Philosophical Traditions_*. Routledge.

Woodfield, A. 1991. Conceptions. *Mind* 100:547-72.

2.7 Meaning Holism

Abbott, B. 2000. Fodor and Lepore on meaning similarity and compositionality. *Journal of Philosophy* 97:454-6.

Becker, K. 1998. On the perfectly general nature of instability in meaning holism. *Journal of Philosophy* 95:635-640.

Bilgrami, A. 1998. Why holism is harmless and necessary. *Philosophical Perspectives* 12:105-126.

Block, N. 1995. An argument for holism. *Proceedings of the Aristotelian Society* 95:151-70.

Uses Putnam's "Ruritania" example to argue that narrow content, if it exists, is holistic. Twins in different communities start with same narrow content, diverge by acquiring new beliefs; so belief change affects narrow content.

Callaway, H.G. 1992. Meaning holism and semantic realism. *Dialectica* 46:41-59.

Churchland, P.M. 1998. Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal Of Philosophy* 95:5-32.

Devitt, M. 1994. A critique of the case for semantic holism. *Philosophical Perspectives* 8:281-306.

Devitt, M. 1994. Semantic localism: Who needs a principled basis? In (R. Casati, B. Smith, & S. White, eds) *_Philosophy and the Cognitive Sciences_*. Holder-Pichler-Tempsky.

Esfeld, M. 1998. Holism and analytic philosophy. *Mind* 107:365-80.

Fodor, J.A. & LePore, E. 1992. *_Holism: A Shopper's Guide_*. Blackwell.
Rebutting arguments for meaning holism: those based on confirmation holism (Quine), normativity of interpretation (Davidson, Dennett, Lewis), and functional-role semantics (Block, Field, Churchland).

Fodor, J.A. & LePore, E. 1993. *Precis of _Holism: A Shopper's Guide_*. *Philosophy and Phenomenological Research* 53:637-682.
A discussion of *_Holism_* with comments by Devitt, Rey, McLaughlin, Brandom, and Churchland, and a reply by Fodor and Lepore.

Gauker, C. 1993. Holism without meaning: A critical review of Fodor and Lepore's *_Holism: A Shopper's Guide_*. *Philosophical Psychology* 6:441-49.

Harrell, M. 1996. Confirmation holism and semantic holism. *Synthese* 109:63-101.

Heal, J. 1994. Semantic holism: Still a good buy. *Proceedings of the Aristotelian Society* 68:325-39.
A critique of Fodor and Lepore. Distinguishes versions of holism, and argues for a weak version. Real thinkers are subjects, which imposes constraints on the interrelations of thoughts. Science fiction is irrelevant here.

Kukla, A. & Kukla, R. 1989. Meaning holism and intentional psychology. *Analysis* 173-53.

Contra Fodor, meaning holism is compatible with intentional psychology. Most psychological generalizations quantify over contents, rather than appealing to specific contents.

Lormand, E. 1996. How to be a meaning holist. *Journal of Philosophy* 93:51-73.

Margolis, E. & Laurence, S. 1998. Multiple meanings and stability of content. *Journal of Philosophy* 5:255-63.

McClamrock, R. 1989. Holism without tears: Local and global effects in cognitive processing. *Philosophy of Science* 56:258-74.

Miller, R.B. 1997. One bad and one not very good argument against holism. *Australasian Journal of Philosophy* 75:234-40.

A nice criticism of Fodor and Lepore's arguments that holism implies (a) the nonexistence of intentional laws and (b) the nonlearnability of language.

Pagin, P. 1997. Is compositionality compatible with holism? *Mind and Language* 12:11-33.

Perry, J. 1994. Fodor and Lepore on holism. *Philosophical Studies* 73:123-58.

The argument from anatomism and the failure of the analytic/synthetic distinction to holism fails. On the many different interpretations of holism and anatomism: there is a reasonable molecularist position.

Senor, T.D. 1992. Two-factor theories, meaning holism, and intentionalistic psychology: A reply to Fodor. *Philosophical Psychology* 5:133-51.

Silverberg, A. 1994. Meaning holism and intentional content. *Pacific Philosophical Quarterly* 75:29-53.

Stich, S.P. 1983. Some evidence against narrow causal theories of belief. In *_From Folk Psychology to Cognitive Science_*. MIT Press.

Talmage, C.J.L. & Mercer, M. 1991. Meaning holism and interpretability. *Philosophical Quarterly* 41:301-15.

Talmage, C.J.L. 1998. Semantic localism and the locality of content. *Erkenntnis* 48:101.

2.8 Mental Content, Misc

Allen, C. 1992. Mental content. *British Journal for the Philosophy of Science* 43:537-53.

Beckermann, A. 1996. Is there a problem about intentionality? *Erkenntnis* 45:1-24.

Bilgrami, A. 1992. *_Belief and Meaning: The Unity and Locality of Mental Content_*. Blackwell.

Bogdan, R.J. 1986. The manufacture of belief. In (R. Bogdan, ed) *_Belief: Form, Content, and Function_*. Oxford University Press.

Butler, K. 1995. Content, context, and compositionality. *Mind and Language* 10:3-24.

Cummins, R. 1991. Form, interpretation, and the uniqueness of content: A response to Morris. *Minds and Machines* 1:31-42.

Morris 1991 is wrong: formal individuation is easy, and objectively determinate content isn't needed. External grounding is also irrelevant.

Devitt, M. 1994. The methodology of naturalistic semantics. *Journal of Philosophy* 91:519-44.

Engel, P. 2000. Wherein lies the normative dimension in meaning and mental

content? Philosophical Studies 100:305-321.

Fodor, J.A. 1987. _Psychosemantics: The Problem of Meaning in the Philosophy of Mind_. MIT Press.

Fodor, J.A. 1990. _A Theory of Content and Other Essays_. MIT Press.

Fodor, J.A. 1994. _The Elm and the Expert_. MIT Press.

Garfield, J.L. 2000. The meanings of "meaning" and "meaning": Dimensions of the sciences of mind. Philosophical Psychology 31:421-440.

Gillett, G. 1992. _Representation, Meaning, and Thought_. Oxford University Press.

Haldane, J.J. 1989. Naturalism and the problem of intentionality. Inquiry 32:305-22.

Haugeland, J. 1990. The Intentionality All-Stars. Philosophical Perspectives 4:383-427.

Intentionality around the diamond: neoCartesianism, neobehaviorism, neopragmatism. 1B=Fodor/Pylyshyn, 2B=Dennett/Quine, 3B=Heidegger/Sellars. SS=Wittgenstein. RF=Searle, CF=Skinner, LF=Rorty/Derrida. Lots of fun.

Horgan, T. 1994. Naturalism and intentionality. Philosophical Studies 76:301-26.

Jacob, P. 1997. _What Minds Can Do: Intentionality in a Non-intentional World_. Cambridge University Press.

Kaye, L.J. 1995. A scientific psychologicistic foundation for theories of meaning. Minds and Machines 5:187-206.

Lehrer, K. 1986. Metamind: Belief, consciousness and intentionality. In (R. Bogdan, ed) _Belief: Form, Content, and Function_. Oxford University Press.

Lycan, W.G. 1986. Thoughts about things. In (M. Brand & R. Harnish, ed) _The Representation of Knowledge and Belief_. University of Arizona Press.

Madell, G. 1989. Physicalism and the content of thought. Inquiry 32:107-21.

Maloney, J.C. 1989. _The Mundane Matter of the Mental Language_. Cambridge University Press.

Martin, C.B. & Pfeifer, K. 1986. Intentionality and the non-psychological. Philosophy and Phenomenological Research 46:531-54.

McDowell, J. 1998. Lecture III: Intentionality as a relation. Journal Of Philosophy 95:471-491.

McGinn, C. 1989. _Mental Content_. Blackwell.

McManus, D. 2000. Boghossian, Miller and Lewis on dispositional theories of meaning. Mind and Language 15:393-399.

Miller, A. 1997. Boghossian on reductive dispositionalism about content: The case strengthened. Mind and Language 12:1-10.

Morris, M. 1991. Why there are no mental representations. Minds and Machines 1:1-30.

There can be no non-stipulative content to non-semantically individuated tokens. Mostly a critique of Cummins; also Fodor and Dennett.

- Newton, N. 1996. *_Foundations of Understanding_*. John Benjamins.
- Pacherie, E. 2000. The content of intentions. *Mind and Language* 15:400-432.
- Peacocke, C. 1986. *_Thoughts: An Essay on Content_*. Blackwell.
- Peacocke, C. 1991. Content and norms in a natural world. In (E. Villanueva, ed) *_Information, Semantics, and Epistemology_*. Blackwell.
- Pollock, J. 1990. Understanding the language of thought. *Philosophical Studies* 58:95-120.
Remarks on a number of aspects of mental content -- narrow, propositional, qualitative -- with respect to functionalism and the language of thought. With comments by Baker.
- Priest, G. 2000. Objects of thought. *Australasian Journal of Philosophy* 78:494-502.
- Prinz, J. 2000. The duality of content. *Philosophical Studies* 100:1-34.
- Schiffer, S. 1981. Truth and the theory of content. In (H. Parret, ed) *_Meaning and Understanding_*. Berlin.
- Schiffer, S. 1987. *_Remnants of Meaning_*. MIT Press.
- Sellars, W. & Chisholm, R. 1957. Intentionality and the mental: A correspondence. *Minnesota Studies in the Philosophy of Science* 2:507-39.
- Stalnaker, R. 1999. *_Context and Content: Essays on Intentionality in Speech and Thought_*. Oxford University Press.
- Sterelny, K. 1990. *_The Representational Theory of Mind_*. Blackwell.
- Stich, S.P. 1982. On the ascription of content. In (A. Woodfield, ed) *_Thought and Object_*. Oxford University Press.
On the tacit theories underlying the folk psychology of belief: beliefs are states associated with typical causal patterns. With remarks on ambiguity, content identity and similarity, and environmental dependence.
- Stich, S.P., and Laurence, S. 1994. Intentionality and naturalism. *Midwest Studies in Philosophy* 19:159-82. Reprinted in (Stich) *_Deconstructing the Mind_*. Oxford University Press, 1996.
Argues that a failure to "naturalize" intentionality won't lead to disasters such as irrationalism, irrelevance, or non-science, whether naturalization is understood as analysis, property identity, supervenience, or whatever.
- Tye, M. 1994. Naturalism and the problem of intentionality. *Midwest Studies in Philosophy* 19:122-42.
There's no deep problem of naturalism about intentionality, as we know it's true already. The real puzzle is that of finding a mechanism to close the gap, e.g. via analysis or essentialism. But naturalism doesn't require that.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 3: Metaphysics of Mind [1026]

=====

3.1 Supervenience [94]

3.1a Psychophysical Supervenience (Kim, etc) [14]

3.1b Supervenience and Physicalism [25]

3.1c Technical Issues in Supervenience [27]

- 3.1d Supervenience, General [28]
- 3.2 Reduction [82]
 - 3.2a Multiple Realizability and Reduction [25]
 - 3.2b Nonreductive Materialism [28]
 - 3.2c Reduction in Psychology [14]
 - 3.2d Miscellaneous [15]
- 3.3 Other Psychophysical Relations [127]
 - 3.3a Physicalism [29]
 - 3.3b Token Identity [7]
 - 3.3c Emergence [58]
 - 3.3d Dualism [21]
 - 3.3e Miscellaneous [12]
- 3.4 Functionalism [63]
 - 3.4a Causal Role Functionalism (Armstrong/Lewis) [17]
 - 3.4b Machine Functionalism (Putnam) [15]
 - 3.4c Miscellaneous [31]
- 3.5 Other Psychophysical Theories [238]
 - 3.5a Logical Behaviorism (Ryle, etc) [53]
 - 3.5b Identity Theory (Smart, Place, etc) [98]
 - 3.5c Eliminative Materialism (Rorty, Feyerabend) [30]
 - 3.5d Anomalous Monism (Davidson) [57]
- 3.6 Mental Causation [58]
- 3.7 Personal Identity [189]
 - 3.7a Personal Identity, General [76]
 - 3.7b Parfit on Personal Identity [49]
 - 3.7c Persons [23]
 - 3.7d Split Brains [22]
 - 3.7e Multiple Personality Disorder [19]
- 3.8 Free Will [129]
- 3.9 The Problem of Other Minds [46]

3.1 Supervenience

3.1a Psychophysical Supervenience (Kim, etc)

Crane, T. 1991. Why indeed? Papineau on supervenience. *Analysis* 51:32-7.
Contra Papineau 1989: the assumption of completeness is false or trivial.
Maybe the mental is part of a complete physics. With response by Papineau.

Elugardo, R. 1988. Against weak psychophysical supervenience. *Dialectica* 42:129-43.
Various objections to Kim's arguments for supervenience. Not all internal states relevant to I/O relations are psychological states. Strange.

Kim, J. 1979. Causality, identity and supervenience in the mind-body problem. *Midwest Studies in Philosophy* 4:31-49.
Supervenience of the mental on the physical is what is required to make mental causation possible. Very nice.

Kim, J. 1982. Psychophysical supervenience. *Philosophical Studies* 41:51-70.
Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).
Internal mental states (i.e. ones that are not rooted outside) supervene on synchronous internal physical states, and internal states are all that is relevant in the explanation of behavior.

Kim, J. 1982. Psychophysical supervenience as a mind-body theory. *Cognition and Brain Theory* 5:129-47.
Distinguishes weak (within-world) vs strong (across-worlds) supervenience.
Relates to reduction, internal/external mental states, and various theories.

Kim, J. 1997. Supervenience, emergence, and realization in the philosophy of mind. In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Lewis, H. 1985. Is the mental supervenient on the physical? In (B. Vermazen & M. Hintikka, eds) *Essays on Davidson*. Oxford University Press.

On some problems with supervenience, the relation between supervenience and reduction, and on reasons for accepting psychophysical supervenience. Loose.

Loar, B. 1993. Can we confirm supervenient properties? In (E. Villanueva, ed) *Naturalism and Normativity*. Ridgeview.

If mental properties are supervenient but irreducible to physical/functional properties, we can't confirm them. Confirmation requires an indispensable explanatory role, which irreducibility precludes. With comments by Schiffer.

Macdonald, C. 1995. Psychophysical supervenience, dependency, and reduction. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Merricks, T. 1998. Against the doctrine of microphysical supervenience. *Mind* 107: 59-72.

Noonan, H. 1999. Microphysical supervenience and consciousness. *Mind* 108:755-9.

Papineau, D. 1989. Why supervenience? *Analysis* 50:66-71.

Psychophysical supervenience follows from completeness of physical laws.

Papineau, D. 1995. Arguments for supervenience and physical realization. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Witmer, D.G. 1998. What is wrong with the manifestability argument for supervenience? *Australasian Journal of Philosophy* 76:84-89.

3.1b Supervenience and Physicalism [see also 3.3a]

Armstrong, D.M. 1982. Metaphysics and supervenience. *Critica* 42:3-17.

Argues that everything is logically supervenient on the physical. Considers classes, possibilities, numbers, universals, and objects of thought.

Bailey, A. 1998. Supervenience and physicalism. *Synthese* 117:53-73.

Chalmers, D.J. 1996. Supervenience and materialism. In *The Conscious Mind* (pp. 41-42). Oxford University Press, 1996.

Charles, D. 1992. Supervenience, composition, and physicalism. In (D. Charles & K. Lennon, eds) *Reduction, Explanation and Realism*. Oxford University Press.

Francescotti, R.M. 1998. Defining "physicalism". *Journal of Mind and Behavior* 19:51-64.

Haugeland, J. 1984. Ontological supervenience. *Southern Journal of Philosophy Supplement* 22:1-12.

Supervenience is all we need for materialism. Various materialist arguments (unity, "nothing but", history, fear of darkness, simplicity, law) don't support physical exhaustion & token identity, over and above supervenience.

Hellman, G. & Thomson, F. 1975. Physicalism: ontology, determination and reduction. *Journal of Philosophy* 72:551-64.

- Hellman, G. & Thomson, F. 1977. Physicalist materialism. *Nous* 11:309-45.
Some applications of the earlier treatment: examples of determination without reduction; the status of properties and universals; the mental; the life sciences; modalities and essentialism; theoretical equivalence.
- Hellman, G. 1985. Determination and logical truth. *Journal of Philosophy* 82:607-16.
Some remarks on determination, physicalism, model theory, and logical truth.
- Horgan, T. 1981. Token physicalism, supervenience, and the generality of physics. *Synthese* 49:395-413.
Argues that the generality of physics should be a supervenience thesis, not token physicalism. Fodor's token physicalism is untenable but might be saved with an appropriate view of events.
- Horgan, T. 1982. Supervenience and microphysics. *Pacific Philosophical Quarterly* 63:29-43.
An account of how all facts supervene on microphysical facts, and how all intrinsic facts supervene on intrinsic microphysical facts.
- Horgan, T. 1984. Supervenience and cosmic hermeneutics. *Southern Journal of Philosophy Supplement* 22:19-38.
Laplacean demon's job: number crunching, plus cosmic hermeneutics to explain high-level truths. All high-level truths follow from low-level by meaning constraints. Application to theoretical/mentalist/everyday terms. Nice.
- Jack, A. 1994. Materialism and supervenience. *Australasian Journal of Philosophy* 72:426-43.
Supervenience is neither necessary nor sufficient for materialism. With various (contentious) counterexamples. So we need a different formulation.
- Kirk, R. 1996. Strict implication, supervenience, and physicalism. *Australasian Journal of Philosophy* 74:244-57.
Argues for strict implication rather than supervenience as a formulation of "minimal physicalism" (unless supervenience is formulated just right).
- Lewis, D. 1983. New work for a theory of universals. *Australasian Journal of Philosophy*.
Formulates a definition of materialism: among worlds where no natural properties alien to our worlds are instantiated, no two differ without differing physically. With a lot of other material on universals.
- Melnyk, A. 1991. Physicalism: From supervenience to elimination. *Philosophy and Phenomenological Research* 51:573-87.
How can supervenience, as a relationship between ontologically distinct properties, be explained? Modal realism and grand-properties don't work. Eliminativism about supervenient properties is the only possibility.
- Moreland, J.P. 1999. Should a naturalist be a supervenient physicalist? *Metaphilosophy* 29:35-57.
- Moser, P. & Trout, J.D. 1996. Physicalism, supervenience, and dependence. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.
- Pettit, P. 1993. A definition of physicalism. *Analysis* 53:213-23.
Defines physicalism in terms of claims that microphysical entities constitute everything and that microphysical laws govern everything. With a reply by Crane.
- Rowlands, M. 1995. *Supervenience and Materialism*. Avebury.
- Seager, W.E. 1988. Weak supervenience and materialism. *Philosophy and*

Phenomenological Research 48:697-709.

Weak supervenience provides a more tenable form of materialism than strong supervenience, because of inverted spectrum possibilities, etc.

Snowdon, P.F. 1989. On formulating materialism and dualism. In (J. Heil, ed) *Cause, Mind, and Reality: Essays Honoring C. B. Martin*. Kluwer.

A construal of materialism in terms of constitution, not identity. Discusses the entailment between physical properties and mental properties; considers a nonreductive physicalism and a primitive dualism.

Sturgeon, S. 1998. Physicalism and overdetermination. *Mind* 107:411-32.

Wilson, J. 1999. How superduper does a physicalist supervenience need to be? *Philosophical Quarterly* 49:33-52.

Witmer, D.G. 1999. Supervenience physicalism and the problem of extras. *Southern Journal of Philosophy* 37:315-31.

3.1c Technical Issues in Supervenience

Bacon, J. 1986. Supervenience, necessary coextensions, and reducibility. *Philosophical Studies* 49:163-76.

A modal-logic analysis of the relations between various notions of supervenience. Most concepts of supervenience entail necessary co-extension, under certain closure assumptions for properties.

Bacon, J. 1995. Weak supervenience supervenes. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Bonevac, D. 1988. Supervenience and ontology. *American Philosophical Quarterly* 25:37-47.

A model-theoretic treatment of supervenience, in terms of relations between theories. Supervenience turns out to be equivalent to reduction.

Bovens, L. 1994. Principles of supervenience. *Australasian Journal of Philosophy* 72:294-301.

Divers, J. 1996. Supervenience for operators. *Synthese* 106:103-12.

Forrest, P. 1988. Supervenience: The grand-property hypothesis. *Australasian Journal of Philosophy* 66:1-12.

A nonreductive supervenience hypothesis: supervenient properties are properties of properties, e.g intrinsic goodness is a property of an object's nature.

Forrest, P. 1992. Universals and universalisability: An interpretation of Oddie's discussion of supervenience. *Australasian Journal of Philosophy* 70:93-98.

Grim, P. 1997. Worlds by supervenience: Some further problems. *Analysis* 2:146-51.

Grimes, T. 1991. Supervenience, determination, and dependency. *Philosophical Studies* 62:81-92.

On dependency supervenience (B properties determine A properties) versus determination supervenience (A properties need B properties).

Grimes, T. 1995. The Tweedledum and Tweedledee of supervenience. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Humberstone, I.L. 1992. Some structural and logical aspects of the notion of

supervenience. Logical Analysis 35:101-37.

Kim, J. 1984. Concepts of supervenience. Philosophy and Phenomenological Research 45:153-76. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Distinguishes weak and strong supervenience. A mistaken proof that strong and global supervenience are equivalent. Strong supervenience implies a kind of reduction, but not an explanatorily useful reduction.

Kim, J. 1987. 'Strong' and 'global' supervenience revisited. Philosophy and Phenomenological Research 48:315-26. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Reasons why global supervenience doesn't entail strong supervenience, and trying to rescue global supervenience as a useful notion. Suggests a similarity-based notion of global supervenience.

Kim, J. 1988. Supervenience for multiple domains. Philosophical Topics 16:129-50. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

How properties in one domain can supervene on properties in another, with or without co-ordination between domains. Relation to global supervenience.

Klagge, J.C. 1995. Supervenience: Model theory or metaphysics? In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Marras, A. 1993. Supervenience and reducibility: An odd couple. Philosophical Quarterly 43:215-222.

Supervenience doesn't entail reducibility, as necessary coextension doesn't suffice, and is incompatible with reducibility, due to ontological asymmetry.

McLaughlin, B.P. 1995. Varieties of supervenience. In (E. Savellos & U. Yalcin, eds) *Supervenience: New Essays*. Cambridge University Press.

Distinguishes modal-operator and possible-worlds versions of supervenience, and explicates global supervenience and its relation to weak and strong. With remarks on multiple-domain supervenience and the relation to reduction.

McLaughlin, B. P. 1997. Supervenience, vagueness, and determination. Philosophical Perspectives 11:209-30.

Melnyk, A. 1997. On the metaphysical utility of claims of global supervenience. Philosophical Studies 87:277-308.

Moser, P.K. 1992. Physicalism and global supervenience. Southern Journal of Philosophy 30:71-82.

Argues that global supervenience has epistemological problems -- how could we ever know that it holds, and that certain worlds are impossible?

Oddie, G. & Tichy, P. 1990. Resplicing properties in the supervenience base. Philosophical Studies 58:259-69.

Closure under resplicing makes supervenience both too narrow and too wide. Weak supervenience is generally too weak to capture the dependence relation.

Oddie, G. 1991. Supervenience and higher-order universals. Australasian Journal of Philosophy 69:20-47.

Paull, R.C. & Sider, T.R. 1992. In defense of global supervenience. Philosophy and Phenomenological Research 52:833-53.

Gives a proof of the distinction between strong and global supervenience that improves on Petrie's, and argues contra Kim that global supervenience is a perfectly reasonable dependence relation for physicalism.

Petrie, B. 1987. Global supervenience and reduction. Philosophy and

Phenomenological Research 48:119-30.

Defending global supervenience: it's weaker than strong supervenience, as base properties of other individuals are relevant. It doesn't entail type or token reducibility. On the relation to implicit definability and reduction.

Post, J.F. 1995. "Global" supervenient determination: Too permissive? In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Shagrir, O. 1999. More on global supervenience. *Philosophy and Phenomenological Research* 59:691-701.

van Cleve, J. 1990. Supervenience and closure. *Philosophical Studies* 58:225-38.

Properties in supervenience relations shouldn't be closed under negation or resplicing, due to bad consequences. With reply by Bacon on resplicing.

3.1d Supervenience, General

Blackburn, S. 1984. Supervenience revisited. In (I. Hacking, ed) *_Exercises in Analysis: Essays by Students of Casimir Lewy_*. Cambridge University Press. On the incompatibility of weak supervenience without strong supervenience and realism. With discussion of various strengths of necessity involved in supervenience claims, and application to moral realism and anomalous monism.

Currie, G. 1984. Individualism and global supervenience. *British Journal for the Philosophy of Science* 35:345-58. How social facts supervene on the totality of individual facts. Application to belief, etc.

Enc, B. 1996. Nonreducible supervenient causation. In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Grimes, T. 1988. The myth of supervenience. *Pacific Philosophical Quarterly* 69:152-60. Supervenience is too weak to function as a dependency relation, as e.g. it can hold in two directions at once.

Hare, R.M. 1984. Supervenience. *Proceedings of the Aristotelian Society* 58:1-16. On the universal conditionals that underlie supervenience, and the necessity thereof. A discussion of the necessity of moral, natural kind, and other sorts of supervenience. Contra Davidson, anomalous supervenience is silly.

Heil, J. 1995. Supervenience redux. In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

Hellman, G. 1992. Supervenience/determination a two-way street? Yes, but one of the ways is the wrong way! *Journal of Philosophy* 89:42-47. Reply to Miller 1990. Miller underestimates the modal force of supervenience and invokes irrelevant dispositional properties.

Horgan, T. 1993. From supervenience to superdupervenience: Meeting the demands of a material world. *Mind* 102:555-86. An overview of supervenience, with focus on the problem of explaining supervenience relations. With remarks on mental causation, emergence, physicalism, and reduction.

Kim, J. 1978. Supervenience and nomological incommensurables. *American Philosophical Quarterly* 15:149-56. Developing and motivating the notion of supervenience. Investigating the relationship to reducibility and definability (equivalence, under certain

conditions), and to microphysical determination.

Kim, J. 1984. Supervenience and supervenient causation. *Southern Journal of Philosophy Supplement* 22:45-56.

On weak/strong supervenience, and high-level causation via supervenience.

Kim, J. 1991. Supervenience as a philosophical concept. *Metaphilosophy* 21:1-27. Reprinted in *_Supervenience and Mind_* (Cambridge University Press, 1993).

A nice overview of supervenience and covariance.

Kim, J. 1993. *_Supervenience and Mind_*. Cambridge University Press.

A collection of articles on supervenience and causation in metaphysics and the philosophy of mind, with some added postscripts.

Kincaid, H. 1987. Supervenience doesn't entail reducibility. *Southern Journal of Philosophy* 25:343-56.

Supervenience doesn't entail reducibility, which is epistemological. The problem's not just huge disjuncts, but also the sharing of bases, no local correlations, and base-properties presupposing supervenient properties.

Kincaid, H. 1988. Supervenience and explanation. *Synthese* 77:251-81.

Argues that lower-level theories can explain supervenient but irreducible higher-level theories, but only under certain conditions, as low-level accounts don't have the relevant kind terms.

Klagge, J.C. 1988. Supervenience: Ontological and ascriptive. *Australasian Journal of Philosophy* 66:461-70.

On supervenience as an ontological relation (via metaphysical necessity) or as an ascriptive relation (via conceptual necessity). The first doesn't preclude the second. Moral realism and mental realism are in the same boat.

Loewer, B. 1995. An argument for strong supervenience. In (E. Savellos & U. Yalcin, eds) *_Supervenience: New Essays_*. Cambridge University Press.

McLaughlin, B.P. 1983. Event supervenience and supervenient causation. *Southern Journal of Philosophy Supplement* 22:71-91.

McLaughlin, B.P. 1994. Varieties of supervenience. In (E. Savellos & O. Yalchin, eds) *_Supervenience: New Essays_*.

On a number of issues: possible worlds vs modal notions, explicating global supervenience, the relation between weak/strong/global supervenience, multiple-domain supervenience, and implications for reduction.

Melnyk, A. 1997. On the metaphysical utility of claims of global supervenience. *Philosophical Studies* 87:277-308.

Merricks, T. 1998. Against the doctrine of microphysical supervenience. *Mind* 107:59-71.

Miller, R.B. 1990. Supervenience is a two-way street. *Journal of Philosophy* 87:695-701.

If supervening properties can make arbitrarily fine distinctions, then physical properties supervene on moral/aesthetic/mental properties.

Noonan, H. 1987. Supervenience. *Philosophical Quarterly* 37:78-85.

Contra Blackburn 1984 on the possibility of weak supervenience without strong supervenience, even with metaphysical necessity; using Nozick's concept structures, or indexical definitions. With application to moral realism.

Post, J.F. 1984. On the determinacy of valuation. *Philosophical Studies* 45:315-33.

Stalnaker, R. 1996. Varieties of supervenience. *Philosophical Perspectives* 10:221-42.

Distinguishes "reductionist" and "metaphysical" conceptions of supervenience. Also discusses the relation between strong and global supervenience, degrees of necessity, and the explanatory role of supervenience.

Teller, P. 1984. The poor man's guide to supervenience and determination. *Southern Journal of Philosophy Supplement* 22:137-62.

Compares the Hellman/Thompson notion of determination with Kim's development of supervenience. Uses these to investigate the concept of materialism, and argues that materialism isn't contingent.

Teller, P. 1985. Is supervenience just disguised reduction? *Southern Journal of Philosophy* 23:93-100.

van Brakel, J. 1996. Interdiscourse or supervenience relations: The primacy of the manifest image. *Synthese* 106:253-97.

Zangwill, N. 1997. Explaining supervenience: Moral and mental. *Journal of Philosophical Research* 22:509-18.

3.2 Reduction

3.2a Reduction and Multiple Realizability

Antony, L.M. & Levine, J. 1997. Reduction with autonomy. *Philosophical Perspectives* 11:83-105.

<

Bechtel, W., & Mundale, J. 1999. Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science* 66:175-207.

Bickle, J. 1992. Multiple realizability and psychophysical reduction. *Behavior and Philosophy* 20:47-58.

Bickle, J. 1995. Connectionism, reduction, and multiple realizability. *Behavior and Philosophy* 23:29-39.

Block, N. 1997. Anti-reductionism slaps back. *Philosophical Perspectives* 11:107-32.

Bolender, J. 1995. Is multiple realizability compatible with antireductionism? *Southern Journal of Philosophy* 33:129-42.

Endicott, R.P. 1991. Macdonald on type reduction via disjunction. *Southern Journal of Philosophy* 29:209-14.

Endicott, R.P. 1989. On physical multiple realization. *Pacific Philosophical Quarterly* 70:212-24.

Endicott, R.P. 1993. Species-specific properties and more narrow reductive strategies. *Erkenntnis* 38:303-21.

On species-specific reductions. These can't reduce standard psychological properties, and problems with intra-species multiple realization can't be circumvented without giving up property reduction for token event identity.

Endicott, R.P. 1998. Collapse of the new wave. *Journal of Philosophy* 95:53-72.

Fodor, J. 1997. Special sciences: Still autonomous after all these years. *Philosophical Perspectives* 11:149-63.

Francescotti, R.M. 1997. What multiple realizability does not show. *Journal of Mind and Behavior* 18:13-28.

The anti-reductionist argument assumes that functional properties aren't physical properties (not even extrinsic physical properties). This, not multiple realizability, does the work.

Heil, J. 1999. Multiple realizability. *American Philosophical Quarterly* 36:189-208.

Kim, J. 1992. Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research* 52:1-26. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

Multiple realization is compatible with reductionism. Jade (= jadeite or nephrite) isn't a scientific kind, and neither are multiply realizable mental properties. So there's no global psychology, just lots of local reductions.

Kistler, M. 1999. Multiple realization, reduction and mental properties. *International Studies in the Philosophy of Science* 13.

Nelson, A. 1985. Physical properties. *Pacific Philosophical Quarterly* 66:268-82.

Some comments on Wilson 1985: some special-science properties may be relevantly different in kind from his expanded physical properties.

Macdonald, C. 1992. Psychological type-type reduction via disjunction. *Southern Journal of Philosophy* 30:65-69.

Mucciolo, L. 1974. The identity thesis and neuropsychology. *Nous* 8:327-42.

Argues contra Fodor and Block that neurological equipotentiality doesn't refute type materialism. Mental states may not be anatomically defined neural states, but they may be more abstract neural holograms.

Nasrin, M. 2000. Multiple realizability: Also a difficulty for functionalism. *Journal of Consciousness Studies* 7:25-34.

Ross, P.A. 1999. The limits of physicalism. *Philosophy of Science* 66:94-116.

Schwartz, J. 1992. Who's afraid of multiple realizability?: Functionalism, reductionism, and connectionism. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

Shagrir, O. 1998. Multiple realization, computation and the taxonomy of psychological states. *Synthese* 114:445-461.

Shapiro, L. 2000. Multiple realizations. *The Journal of Philosophy* 97:635-654.

Wilson, M. 1985. What is this thing called 'pain'? -- The philosophy of science behind the contemporary debate. *Pacific Philosophical Quarterly* 66:227-67.

Argues for type-type identities and for an expanded view of the physical, as properties from physics exhibit the same sort of multiple realizability as functional properties. Sophisticated, with many interesting examples.

Zangwill, N. 1995. Supervenience, reduction, and infinite disjunction. *Philosophia* 24:321-30.

3.2b Nonreductive Materialism [see also 3.5d]

Barrett, J. 1995. Causal relevance and nonreductive physicalism. *Erkenntnis* 42:339-62.

Beckermann, A. 1992. Reductive and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for*

Nonreductive Physicalism_. De Gruyter.

On varieties of physicalism with respect to reduction: semantic physicalism, identity theory, supervenience, and the denial of emergence. Advocates a version on which physical states realize mental states.

Beckermann, A, Flohr, H. & Kim, J. (eds) 1992. _Emergence or Reduction?: Prospects for Nonreductive Physicalism_. De Gruyter.

Boyd, R. 1980. Materialism without reductionism: What physicalism does not entail. In (N. Block, ed) _Readings in the Philosophy of Psychology_, Vol 1. MIT Press.

Dupre, J. 1988. Materialism, physicalism, and scientism. Philosophical Topics 16:31-56.

Arguing for a pluralistic conception. With criticism of Churchland's reductionism, Davidson's token identity, and more generally reverential "scientism". Reductionist explanation is not the general rule.

Ellis, R. 2000. Consciousness, self-organization, and the process-substratum relation: Rethinking nonreductive physicalism. Philosophical Psychology 13:173-190.

Fodor, J.A. 1974. Special sciences. Synthese 28:97-115. Reprinted in (N. Block, ed) _Readings in the Philosophy of Psychology_ (MIT Press, 1980).

Psychological kinds can't be reduced to physical kinds, due to cross-classification, although token physicalism still holds. How to maintain the generality of physics without a reductionist unity of science.

Francescotti, R.M. 1998. The nonreductionist's troubles with supervenience. Philosophical Studies 89:105-24.

Horgan, T. 1993. Nonreductive materialism and the explanatory autonomy of psychology. In (S. Wagner & R. Warner, eds) _Naturalism: A Critical Appraisal_. University of Notre Dame Press.

Gives four constraints on interlevel connections, and some arguments against reductionism and for the autonomy of psychology. Argues that supervenience fact are themselves in need of explanation.

Horgan, T. 1994. Nonreductive materialism. In (R. Warner & T. Szubka, eds) _The Mind-Body Problem: A Guide to the Current Debate_. Blackwell.

Kernohan, A. 1988. Non-reductive materialism and the spectrum of mind-body identity theories. Dialogue 27:475-88.

Classifying psychophysical theories by the status (necessary, lawful, anomalous, false) of psychophysical/psychological generalizations. Defending autonomous monism: nonreductive materialism with psychological laws.

Kim, J. 1989. The myth of non-reductive materialism. Proceedings and Addresses of the American Philosophical Association 63(3):31-47. Reprinted in _Supervenience and Mind_ (Cambridge University Press, 1993).

Somewhat loose arguments that non-reductive physicalist realism is untenable. Anomalous monism makes the mental irrelevant, functionalism is compatible with species-specific reduction, and supervenience is weak or reductive.

Kim, J. 1992. The nonreductivist's trouble with mental causation. In (J. Heil & A. Mele, eds) _Mental Causation_. Oxford University Press. Reprinted in _Supervenience and Mind_ (Cambridge University Press, 1993).

Argues that nonreductive materialism implies downward causation (as the mental has more causal powers than the physical alone), and that downward causation violates the causal closure of the physical.

Kim, J. 1992. "Downward causation" in emergentism and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) _Emergence or

Reduction?: Prospects for Nonreductive Physicalism_. De Gruyter.

Argues that nonreductive materialism is just like 1930s emergentism, with the the mental contributing new causal powers, and so implies downward causation.

Kirk, R. 1996. How physicalists can avoid reductionism. Synthese 108:157-70.

Contra Kim, physicalists can avoid reduction by embracing strict implication.

Margolis, J. 1978. _Persons and Minds: The Prospects of Non-Reductive Materialism_. D. Reidel.

Marras, A. 1993. Psychophysical supervenience and nonreductive materialism. Synthese 95:275-304.

Marras, A. 1994. Nonreductive materialism and mental causation. Canadian Journal of Philosophy 24:465-93.

Melnyk, A. 1995. Two cheers for reductionism, or, the dim prospects for nonreductive materialism. Philosophy of Science 62:370-88.

Melnyk, A. 1998. The prospects for Kirk's nonreductive physicalism. Australasian Journal of Philosophy 76:323-32.

Loar, B. 1992. Elimination versus nonreductive physicalism. In (D. Charles & K. Lennon, eds) _Reduction, Explanation and Realism_. Oxford University Press.

Papineau, D. 1992. Irreducibility and teleology. In (D. Charles & K. Lennon, eds) _Reduction, Explanation and Realism_. Oxford University Press.
Non-reductive physicalism is a mystery unless we invoke teleology.

Pereboom, D. & Kornblith, H. 1991. The metaphysics of irreducibility. Philosophical Studies 63:125-45.

Explicating anti-reductionism: mental causal powers are constituted of physical causal powers, but aren't type- or token-identical to them. Against arguments from local reduction, neuroscience, explanatory exclusion, etc.

Silvers, S. 1997. Nonreductive naturalism. Theoria 12:163-84.

Smith, A.D. 1993. Non-reductive physicalism? In (H. Robinson, ed) _Objections to Physicalism_. Oxford University Press.

A careful discussion of how to characterize physicalism, in terms of identity or supervenience, and argues that physicalism must reduce (bowdlerize) qualia to something they are not, as physicalism requires topic-neutral analyses.

Ten Elshof, G. 1997. Supervenient difficulties with nonreductive physicalism: A critical analysis of supervenience physicalism. Kinesis 24:3-22.

van Gulick, R. 1992. Nonreductive materialism and the nature of intertheoretical constraint. In (A. Beckermann, H. Flohr, & J. Kim, eds) _Emergence or Reduction?: Prospects for Nonreductive Physicalism_. De Gruyter.
On how a nonreductive materialism can handle problems about mental causation, psychophysical dependencies, and qualia. A teleofunctionalist view with different conceptual frameworks, but mental properties physically realized.

Wedgwood, R. 2000. The price of non-reductive physicalism. Nous 34:400-421.

3.2c Reduction in Psychology

Bickle, J. 1995. Psychoneural reduction of the genuinely cognitive: Some accomplished facts. Philosophical Psychology 8:265-85.

Argues that cognitive theories have already been reduced to neurobiology in some domains, such as associative learning.

Churchland, P.M. 1982. Is 'thinker' a natural kind? *Dialogue* 21:223-38.

Psychology shouldn't be autonomous from natural science. By analogy with biology, nature provides (a) conceptual insight, and (b) real constraints, e.g. thermodynamic ones. Biology and psychology are continuous.

Gaito, J. 1960. Description, explanation, and reductionism in psychology. *Psychological Reports* 6:203-5.

Gaito, J. & Leonard, D. 1965. Philosophical and empirical reductionism in psychology. *Journal of General Psychology* 72:69-75.

Hardcastle, V.G. 1992. Reduction, explanatory extension, and the mind/brain sciences. *Philosophy of Science* 59:408-28.

The relationship between psychology and neuroscience is best characterized not by reduction but by explanatory extension, where each field is enriched by the other. With a number of examples from recent empirical work.

Hyland, M.E. 1995. Against nomological reductionism in psychology: A response to Robinson. *New Ideas in Psychology* 13:9-11.

Jessor, R. 1958. The problem of reductionism in psychology. *Psychological Review* 65:170-78.

Marras, A. 1990. Reduction in psychology. *Acta Analytica* 6:65-78.

Martindale, R.L. & Seidel, R.J. 1959. Reductionism: Its prodigal encores. *Psychological Reports* 5:213-16.

Montgomery, R. 1990. The reductionist ideal in cognitive psychology. *Synthese* 85:279-314.

Anti-reductionism needn't be ad hoc (contra Churchland). Although evolution provides some pressure for 1-1 psychophysical mappings, there are significant countervailing forces, e.g. in vision, memory, learning, and language use.

Olszewsky, T.M. 1975. Dispositions and reductionism in psychology. *Journal for the Theory of Social Behavior* 5:129-44.

Putnam, H. 1974. Reductionism and the nature of psychology. *Cognition* 2:131-46.

Richardson, R.C. 1999. Cognitive science and neuroscience: New wave reductionism. *Philosophical Psychology* 12:297-307.

Sloane, E.H. 1945. Reductionism. *Psychological Review* 52:214-23.

3.2d Reduction, Misc

Beckermann, A. 1997. Property physicalism, reduction, and realization. In (M. Carrier & P. Machamer, eds) *Mindscales: Philosophy, Science, and the Mind*. Pittsburgh University Press.

Bickle, J. 1996. New wave psychophysical reductionism and the methodological caveats. *Philosophy and Phenomenological Research* 56:57-78.

Bickle, J. 1997. *Psychoneural Reductionism: The New Wave*. MIT Press.

Brooks, D.H.M. 1994. How to perform a reduction. *Philosophy and Phenomenological Research* 54:803-14.

Reduction comes to supervenience plus explicability. Thus biconditionals, multiple realizability, etc, are irrelevant. Biology is already reduced (mostly via functional explanation), and psychology looks promising. Nice.

Bunzl, M. 1987. Reductionism and the mental. *American Philosophical Quarterly*

24:181-9.

On the links between supervenience, reduction, and explanation. Supervenience is compatible with reductive explanation of a localized variety. We don't need laws, but explanatory links.

Causey, R.L. 1972. Attribute identities in microreductions. *Journal of Philosophy* 69:407-22.

Combes, R. 1988. Ockhamite reductionism. *International Philosophical Quarterly* 28:325-36.

Foss, J. 1995. Materialism, reduction, replacement, and the place of consciousness in science. *Journal of Philosophy* 92:401-29.

Hill, C.S. 1984. In defense of type materialism. *Synthese* 59:295-320.

Kitcher, P.S. 1980. How to reduce a functional psychology. *Philosophy of Science* 47:134-40.

Contra Richardson 1979, a purely functional psychology is irreducible. The genetics analogy is misleading; multiple realizations can't explain high-level laws.

Papineau, D. 1985. Social facts and psychological facts. In (G. Currie & A. Musgrave, eds) *Popper and the Human Sciences*. Martinus Nijhoff.

Mind is not reducible to body, but societies reduce to individuals. Multiple realization is in tension with predictability. Natural selection resolves the tension for the mental, but cannot for the social.

Richardson, R.C. 1979. Functionalism and reductionism. *Philosophy of Science* 46:533-58.

Argues that functionalism is compatible with reductionism, by analogies. Genetics has multiple realization and multiple function; reduction doesn't require biconditionals. With remarks on the de facto autonomy of psychology.

Richardson, R.C. 1982. How not to reduce a functional psychology. *Philosophy of Science* 49:125-37.

Response to Kitcher 1980. Reductions are usually domain-specific, and high-level regularities are indeed explained.

Sarkar, S. 1992. Models of reduction and categories of reductionism. *Synthese* 91:167-94.

Wimsatt, W. 1976. Reductionism, levels of organization, and the mind-body problem. In (G. Globus, ed) *Consciousness and the Brain*. Plenum Press.

Excellent coverage of the notion of level and its applicability to mind.

3.3 Other Psychophysical Relations

3.3a Physicalism [see also 1.3, 1.7b, 3.1b, 3.2b, 3.5]

Crane, T. 1991. All God has to do. *Analysis* 51:235-44.

If there are no contingent psychophysical laws, then there are no mental properties. So physicalism/supervenience is false; God had extra work to do.

Crane, T. 1993. A definition of physicalism: Reply to Pettit. *Analysis* 53:224-27.

Crane, T. & Mellor, D.H. 1990. There is no question of physicalism. *Mind* 99:185-206.

Physical sciences have no ontological authority over the mental. Considers and dismisses arguments from laws, causation, reduction, supervenience.

Daly, C. 1995. Does physicalism need fixing? *Analysis* 55:135-41.

Francescotti, R. 2000. Ontological physicalism and property pluralism: Why they are incompatible. *Pacific Philosophical Quarterly* 81:349-362.

Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, & S. White, eds) *Philosophy and the Cognitive Sciences*.

Holder-Pichler-Tempsky.

On why materialism requires conceptual analysis to locate mental properties in the natural world. Even a posteriori necessary connections have to be backed by a priori links. With remarks on supervenience. A nice paper.

Kirk, R. 1979. From physical explicability to full-blooded materialism. *Philosophical Quarterly* 29:229-37.

If every physical events has a physical explanation, and the mental is causally efficacious, then mental facts are strictly implied by physical facts. A nice argument.

Kirk, R. 1982. Physicalism, identity, and strict implication. *Ratio* 24:131-41.

Materialism doesn't need a identity thesis. The requirement that mental facts are entailed by physical facts plays the role played by Kripke's requirement of necessary identity, and is more reasonable.

Kirk, R. 1996. Physicalism lives. *Ratio* 9:85-89.

Nothing in the arguments of Crane and Mellor 1990 count against a physicalism based on strict implication.

Madell, G. 1988. *Mind and Materialism*. Edinburgh University Press.

On the problems posed for materialism by intentionality, autonomy, awareness, and indexicality. Tentatively advocates a Cartesian position.

McGinn, C. 1980. Philosophical materialism. *Synthese* 44:173-206. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Melnyk, A. 1994. Being a physicalist: How and (more importantly) why.

Advocates "realization physicalism": all properties are either physical or functional properties realized by physical ones. This achieves unity between sciences better than alternatives, and avoids overdetermination.

Melnyk, A. 1996. Formulating physicalism: Two suggestions. *Synthese* 105:381-407.

Discusses two formulations of physicalism: requiring high-level properties to be disjunctions of physical states, or to be functional properties realized physically. Tentatively endorses the latter.

Melnyk, A. 1997. How to keep the 'physical' in physicalism. *Journal of Philosophy* 94:622-637.

Montero, B. 1999. The body problem. *Nous* 33:183-200.

Moser, P.K. 1996. Physicalism and mental causes: Contra Papineau. *Analysis* 56:263-67.

Nagel, E. 1949. Are naturalists materialists? *Journal of Philosophy* 42:515-53.

Papineau, D. 1994. *Philosophical Naturalism*. Blackwell.

Pettit, P. 1993. A definition of physicalism. *Analysis* 53:213-23.

Physicalism is the claim that (1) There are microphysical entities, (2) Microphysical entities constitute everything, (3) There are microphysical

regularities, (4) Microphysical regularities govern everything.

Pettit, P. 1994. Microphysicalism without contingent micro-macro laws. *Analysis* 54:253-57.

Pettit, P. 1995. Microphysicalism, dottism, and reduction. *Analysis* 55:141-46.

Poland, J. 1994. *_Physicalism: The Empirical Foundations_*. Oxford University Press.

Ravenscroft, I. 1997. Physical properties. *Southern Journal Of Philosophy* 35:419-431.

Robinson, D. 1991. On Crane and Mellor's argument against physicalism. *Mind* 100:135-36.

Robinson, H. (ed) 1993. *_Objections to Physicalism_*. Oxford University Press.

Sheldon, W.H. 1946. Are naturalists materialists? *Journal of Philosophy* 43:197-209.

Sober, E. 1999. Physicalism from a probabilistic point of view. *Philosophical Studies* 95:135-74.

Sturgeon, S. 1998. Physicalism and overdetermination. *Mind* 107:411-432.

Wilkes, K.V. 1973. *_Physicalism_*. Routledge and Kegan Paul.

3.3b Token Identity [see also 3.5b, 3.5d]

Foster, J. 1994. The token-identity thesis. In (R. Warner & T. Szubka, eds) *_The Mind-Body Problem: A Guide to the Current Debate_*. Blackwell.

Horgan, T. & Tye, M. 1985. Against the token identity theory. In (B. McLaughlin & E. LePore, eds) *_Action and Events_*. Blackwell.

We individuate mental events by their causal role, but we can't individuate causes uniquely. So each mental event has multiple physical correlates, and token identity doesn't hold.

Hornsby, J. 1981. Which physical events are mental events? *Proceedings of the Aristotelian Society* 55:73-92.

Haugeland, J. 1982. Weak supervenience. *American Philosophical Quarterly* 19:93-103.

Supervenience doesn't imply token identity, and Davidson's argument for token identity equivocates on "event". But weak supervenience (mentally discernible worlds are physically discernible) is all we need. With nice examples.

Leder, D. 1985. Troubles with token identity. *Philosophical Studies* 47:79-94. Physical/psychological token identity is no good: you can't individuate physical events without psychological predicates.

Lurie, Y. 1978. Correlating brain states with psychological phenomena. *Australasian Journal of Philosophy* 56:135-44.

Can't isolate the physical token of a belief, say, as it's always accompanied by other beliefs. Meaning doesn't come in discrete tokens.

Peacocke, C. 1979. Argument for token identity. In *_Holistic Explanation_*. Oxford University Press.

3.3c Emergence

- Alexander, S. 1920. *Space, Time, and Deity*. Macmillan.
- Atkin A. 1992. On consciousness: What is the role of emergence? *Medical Hypotheses* 38:311-14.
- Beckermann, A. 1992. Supervenience, emergence, and reduction. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.
On varieties of supervenience and of emergence, and of what is required for reduction. Argues that reduction involves general explanatory connections, whereas emergence involves unique and ultimate bridge laws.
- Beckermann, A, Flohr, H. & Kim, J. (eds) 1992. *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.
- Bedau, M. 1997. Weak emergence. *Philosophical Perspectives* 11:375-399.
- Berenda, C.W. 1953. On emergence and prediction. *Journal of Philosophy* 50:269-74.
- Bergmann, G. 1944. Holism, historicism, and emergence. *Philosophy of Science* 11:209-21.
- Broad, C.D. 1925. *The Mind and its Place in Nature*. Routledge and Kegan Paul.
- Bruntrup, G. 1998. Is psychophysical emergentism committed to dualism? The causal efficacy of emergent mental properties. *Erkenntnis* 48:133-51.
- Bunge, M. 1977. Emergence and the mind. *Neuroscience* 2:501-9.
- Garnett, A.C. 1942. Scientific method and the concept of emergence. *Journal of Philosophy* 39:477-86.
- Haldane, J. 1996. The mystery of emergence. *Proceedings of the Aristotelian Society* 96:261-67.
A defence of radical emergence against Spencer-Smith 1995.
- Hasker, W. 1982. Emergentism. *Religious Studies* 18:473-88.
- Hasker, W. 1999. *The Emergent Self*. Cornell University Press.
- Henle, P. 1942. The status of emergence. *Journal of Philosophy* 39:486-93.
- Humphreys, P. 1996. Aspects of emergence. *Philosophical Topics* 24:53-71.
- Humphreys, P. 1997. How properties emerge. *Philosophy of Science* 64:1-17.
- Humphreys, P. 1997. Emergence, not supervenience. *Philosophy of Science Supplement* 64:337-45.
- Jones, D.H. 1972. Emergent properties, persons, and the mind-body problem. *Southern Journal of Philosophy* 10:423-33.
- Kekes, J. 1966. Physicalism, the identity theory, and the concept of emergence. *Philosophy of Science* 33:360-75.
- Kim, J. 1999. Making sense of emergence. *Philosophical Studies* 95:3-36.
- Klee, J. 1984. Microdeterminism and concepts of emergence. *Philosophy of Science* 51:44-63.
- Lovejoy, A.O. 1927. The meanings of "emergence" and its modes. In (E.S.

Brightman, ed) *_Proceedings of the Sixth International Congress of Philosophy_*. Longmans, Green, and Co.

Lowe, E. J. 2000. Causal closure principles and emergentism. *Philosophy* 75:571-586.

Lowry, A. 1974. A note on emergence. *Mind* 83:276-77.

Mackenzie, W.L. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:56-68.

Margolis, J. 1986. Emergence. *Philosophical Forum* 17:271-95.

McLaughlin, B.P. 1992. The rise and fall of British emergentism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.

A careful account of British emergentism. Explicates their view of emergent causal powers and laws in terms of fundamental configurational forces, a coherent idea that turned out to be false. An excellent paper.

Meehl, P.E. & Sellars, W. 1956. The concept of emergence. In (H. Feigl & M. Scriven, eds) *_Minnesota Studies in the Philosophy of Science_*, vol. 1. University of Minnesota Press.

Morgan, C.L. 1923. *_Emergent Evolution_*. Williams and Norgate.

Morris, C.R. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:49-55.

Newman, D. 1996. Emergence and strange attractors. *Philosophy of Science* 63:245-61.

Newman, D.V. 2001. Chaos, emergence, and the mind-body problem. *Australasian Journal of Philosophy* 79:180-96.

O'Connor, T. 1994. Emergent properties. *American Philosophical Quarterly* 31:91-104.

Argues against Alexander's and van Cleve's accounts of emergence, instead suggesting an account in terms of supervenience, non-structurality, and downward causation.

Pap, A. 1951. The concept of absolute emergence. *British Journal for the Philosophy of Science* 2:302-11.

Pepper, S.C. 1926. Emergence. *Journal of Philosophy* 23:241-45.

Peters, S.L. 1995. *_Emergent Materialism: A Proposed Solution to the Mind-Body Problem_*. University Press of America.

Pihlstrom, S. 1999. What shall we do with emergence? A survey of a fundamenta; issue in the metaphysics and epistemology of science. *South African Journal of Philosophy* 18:192-210.

Pluhar, E. 1978. Emergence and reduction. *Studies in History and Philosophy of Science* 9:279-89.

Ripley, C. 1984. Sperry's concept of consciousness. *Inquiry* 27:399-423.
An in-depth analysis of Sperry's views on consciousness. Sperry is not a dualist; he believes in "structural causation" based on emergent properties.

Rohrlich, F. 1997. Cognitive emergence. *Philosophy of Science Supplement* 64:346-58.

- Rueger, A. 2000. Robust supervenience and emergence. *Philosophy of Science* 67:466-491.
- Russell, E.S. 1926. The notion of emergence. *Aristotelian Society Supplement* 6:39-48.
- Schroder, J. 1998. Emergence: Non-deducibility or downwards causation? *Philosophical Quarterly* 48:433-52.
- Silberstein, M. 1998. Emergence and the mind-body problem. *Journal of Consciousness Studies* 5:464-82.
- Silberstein, M. & McGeever, J. 1999. The search for ontological emergence. *Philosophical Quarterly* 49:182-200.
- Smart, J.J.C. 1981. Physicalism and emergence. *Neuroscience* 6:109-13.
- Spencer-Smith, R. 1995. Reductionism and emergent properties. *Proceedings of the Aristotelian Society* 95:113-29.
Distinguishes radical, epistemic, and interactional emergence, favoring the latter. With consideration of qualia as a radical emergent.
- Sperry, R.W. 1969. A modified concept of consciousness. *Psychological Review* 76:532-36.
Consciousness is an emergent property of brain dynamics that itself governs low-level flow of excitation. Midway between mentalism and materialism.
- Sperry, R.W. 1991. In defense of mentalism and emergent interaction. *Journal of Mind and Behavior* 12:221-245.
- Stace, W.T. 1939. Novelty, indeterminism, and emergence. *Philosophical Review* 48:296-310.
- Stephan, A. 1992. Emergence -- a systematic look at its historical facets. In (A. Beckermann, H. Flohr, & J. Kim, eds) *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.
On different ways of understanding emergence: as nonadditivity, novelty, nonpredictability, nondeducibility; and on problems about qualia and downward causation.
- Stephan, A. 1997. Armchair arguments against emergence. *Erkenntnis* 46:305-14.
- Teller, P. 1992. A contemporary look at emergence. In (A. Beckermann, H. Flohr, & J. Kim, eds) *_Emergence or Reduction?: Prospects for Nonreductive Physicalism_*. De Gruyter.
An attempt to explicate "emergent" properties in terms of relational properties. Argues that even problem cases, e.g. space-time separation and phenomenal properties, might be treated this way.
- van Cleve, J. 1990. Mind -- dust or magic? Panpsychism versus emergence. *Philosophical Perspectives* 4:215-226.
On Nagel 1979: emergence is more plausible than panpsychism. A construal of emergence as nomological supervenience without logical supervenience.
- vandervert, L.R. 1991. On the modeling of emergent interaction: Which will it be, the laws of thermodynamics or Sperry's "wheel" in the subcircuitry? *Journal of Mind and Behavior* 12:535-39.
- Wimsatt, W.C. 1997. Aggregativity: Reductive heuristics for finding emergence. *Philosophy of Science* 64:372-84.
- Wynn, M. 1999. Emergent phenomena and theistic explanation. *International Philosophical Quarterly* 39:141-55.

3.3d Dualism [see also 1.3f, 1.4e, 1.4f]

Averill, E.W. & Keating, B. 1981. Does interactionism violate a law of classical physics? *Mind* 90:102-7.

Interactionism is compatible with conservation of energy and momentum: the mind exerts a non-physical force on the brain.

Bricke, J. 1975. Interaction and physiology. *Mind* 84:255-9.

Efron, A. 1992. Residual asymmetric dualism: A theory of mind-body relations. *Journal of Mind and Behavior* 13:113-36.

Evans, S. 1981. Separable souls: A defense of minimal dualism. *Southern Journal of Philosophy* 19.

Herbert, R.T. 1998. Dualism/materialism. *Philosophical Quarterly* 48:159-75.

Larmer, R. 1986. Mind-body interactionism and the conservation of energy. *International Philosophical Quarterly* 26:277-85.

Various arguments about interactionism based on conservation of energy. C of E only applies to causally isolated systems, so objections beg the question.

Lowe, E.J. 1992. The problem of psychophysical causation. *Australasian Journal of Philosophy* 70:263-76.

Argues that there can be interaction without breaking physical laws: e.g. by basic psychic forces, or by varying physical constants, or especially by arranging fractal trees of physical causation leading to behavior.

Lowe, E.J. 1993. The causal autonomy of the mental. *Mind* 102:629-44.

Mills, E. 1996. Interactionism and overdetermination. *American Philosophical Quarterly* 33:105-115.

Argues that interactionist dualism is compatible with the causal closure of the physical, if we allow causal overdetermination; and there is a strong case for the latter.

Mills, E. 1997. Interactionism and physicality. *Ratio* 10:169-83.

O'Leary-Hawthorne, J. & McDonough, J.K. 1998. Numbers, minds, and bodies: A fresh look at mind-body dualism. *Philosophical Perspectives* 12:349-371.

Pap, A. 1952. Semantic analysis and psychophysical dualism. *Mind*.

Pietroski, P.M. 1994. Mental causation for dualists. *Mind and Language* 9:336-66.

Popper, K.R. 1953. Language and the body-mind problem: A restatement of interactionism. In *Proceedings of the 11th International Congress of Philosophy*. Reprinted in *Conjectures and the Growth of Scientific Knowledge*. Basic Books, 1962.

Popper, K.R. 1955. A note on the body-mind problem. *Analysis* 15:131-35.

Popper, K.R. 1977. Natural selection and the emergence of mind.

Scheffler, I. 1950. The new dualism: Psychological and physical terms. *Journal of Philosophy*.

Sellars, W. 1954. A note on Popper's argument for dualism. *Analysis* 15:23-24.

Sussman, A. 1981. Reflection on the chances for a scientific dualism. *Journal of Philosophy* 78:95-118.

Dualism is an empty hypothesis. Everything must be matter, though we may have to expand the notion of matter.

Richardson, R.C. 1982. The 'scandal' of Cartesian dualism. *Mind* 91:20-37.

van Rooijen, K. 1987. Interactionism and evolution: A critique of Popper. *British Journal for the Philosophy of Science* 38:87-92.

3.3e Psychophysical Relations, Misc

Campbell, K. 1983. Abstract particulars and the philosophy of mind. *Australasian Journal of Philosophy* 61:129-41.

Caston, V. 1997. Epiphenomenalisms ancient and modern. *Philosophical Review* 106:309-363.

Hedman, C.G. 1970. On correlating brain states with psychological states. *Australasian Journal of Philosophy* 48:247-51.

Heil, J. 1992. *The Nature of True Minds*. Cambridge University Press.

Honderich, T. 1981. Psychophysical law-like connections and their problems. *Inquiry* 24:277-303.

Defending lawlike connections between physical states & conscious occurrents. Contra anomalous monism and identity theory for occurrents. But occurrents may not be causally efficacious. Comments by Wilson/Sprigge/Mackie/Stich.

McGinn, C. 1978. Mental states, natural kinds and psychophysical laws. *Proceedings of the Aristotelian Society* 52:195-220. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Argues that mental kinds are not natural kinds, and don't have real essences but nominal essences. For this reason, there are no psychophysical laws. With remarks on psychological laws, and the role of behavior.

Schechtman, M. 1997. The brain/body problem. *Philosophical Psychology* 10:149-64.

Scheerer, E. 1994. Psychoneural isomorphism: Historical background and current relevance. *Philosophical Psychology* 7:183-210.

Skillen, A. 1984. Mind and matter: a problem which refuses dissolution. *Mind* 93:514-26.

Physical completeness, mental causation, non-reductionism are inconsistent. Ryle and Putnam are closet dualists, and Davidson's an epiphenomenalist.

Steward, H. 1997. *The Ontology of Mind: Events, Processes, and States*. Oxford University Press.

Tye, M. 1989. *The Metaphysics of Mind*. Cambridge University Press.

van Gelder, T. 1998. Monism, dualism, pluralism. *Mind and Language* 13:76-97.

3.4 Functionalism [see also 1.8, 4.6]

3.4a Causal Role Functionalism (Armstrong/Lewis)

Armstrong, D.M. 1968. *A Materialist Theory of the Mind*. Routledge and Kegan Paul.

Mental states should be analyzed as states that are apt to bring about certain kinds of behavior. Analysis of all kinds of mental states as such. With comments on dualism, behaviorism, identity theory, and consciousness.

Armstrong, D.M. 1970. The nature of mind. In (C. Borst, ed) *The Mind/Brain Identity Theory*. Macmillan. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Mental states are internal states that are apt to cause certain behaviors. A synthesis between the "thesis" of idealism and the "antithesis" of behaviorism. With defense against objections from consciousness.

Clark, A. 1986. Psychofunctionalism and chauvinism. *Philosophy of Science* 53:535-59.

Psychofunctionalism can evade chauvinism by specifying different functional identifications within each species. Applying same mental terms to each is justified by theory similarity; but it still isn't analytic functionalism.

Goldstein, I. 1994. Identifying mental states: A celebrated hypothesis refuted. *Australasian Journal of Philosophy* 72:46-62.

Against functionalism: experiences have intrinsic introspectible acausal properties, such as duration, felt location, and unpleasantness. Both analytic and empirical functionalism fail.

Horgan, T. 1984. Functionalism and token physicalism. *Synthese* 59:321-38.

Formalizing versions of functionalism, and seeing which entail token physicalism and/or type physicalism. On the most plausible versions, we have token physicalism without type physicalism.

Hornsby, J. 1984. On functionalism, and on Jackson, Pargetter, and Prior on functionalism. *Philosophical Studies* 46:75-96.

Jackson, F., Pargetter, R. & Prior, E.W. 1982. Functionalism and type-type identity theories. *Philosophical Studies* 42:209-25.

Functionalism is compatible with type identity, as e.g. "pain" designates the state-type that fills the right functional role in an organism at a given time, i.e. a brain state. Contra Kripke, pain is not a rigid designator.

Kernohan, A. 1990. Lewis's functionalism and reductive materialism. *Philosophical Psychology* 3:235-46.

Argues that Lewis's functionalism founders on the specification of behavior. Described intentionally => non-materialist; physically => chauvinist.

Lewis, D. 1966. An argument for the identity theory. *Journal of Philosophy* 63:17-25. Reprinted in *Philosophical Papers, Vol. 1* (Oxford University Press, 1980).

Causal roles are definitive of mental states. Since physical states fill these causal roles (by the explanatory adequacy of physics), mental states are physical states.

Lewis, D. 1972. Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50:249-58. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Mental states can be defined, via a Ramsey-sentence analysis of the platitudes of folk psychology, as entities that fill causal roles specified by the analysis. These fillers turn out to be physical.

Lewis, D. 1978. Mad pain and martian pain. In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol. 1. MIT Press.

Accounting for both pains that don't play the usual causal role and for pains that are realized in different substances, by a mixed theory: pain is the physical state that typically occupies a certain causal role in a population.

McGinn, C. 1980. Functionalism and phenomenalism: A critical note.

Australasian Journal of Philosophy 58:35-46. Reprinted in *The Problem of Consciousness* (Blackwell, 1991).

Functionalism (reducing the mental to its effects on the physical) is no more plausible than phenomenalism (reducing the physical to its effects on the mental).

Owens, J. 1982. The failure of Lewis's functionalism. *Philosophical Quarterly* 36:159-73.

Lewis's original theory leads to Kripkean reference-fixing, so chauvinism. Token functionalism can't deal with paralytics. Species-relative functionalism fails as pain is intrinsic, not extrinsic.

Rogler, E. 2000. On David Lewis' philosophy of mind. *Protosociology* 14:285-311.

Sayward, C. 1995. Taking actions seriously. *Behavior and Philosophy* 23:51-60.

Shoemaker, S. 1981. Some varieties of functionalism. *Philosophical Topics* 12:93-119. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Fleshing out Ramsey-sentence functionalism; against Lewis's "mad pain" mixed theory; relating functionalism to the causal theory of properties. Empirical functionalism is chauvinistic so probably false. A terrific, in-depth paper.

Tye, M. 1983. Functionalism and type physicalism. *Philosophical Studies* 44:161-74.

Contra Lewis: Functionalism isn't compatible with type physicalism. There are intra-population difficulties with species-relative construals, and individual-relative construals can still have multiple fillers.

3.4b Machine Functionalism (Putnam) [see also 4.8]

Putnam, H. 1960. Minds and machines. In (S. Hook, ed) *Dimensions of Mind*. New York University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

The relationship between mental and physical states is just like that between logical and structural states of Turing Machines, so no great mystery. With comments on privacy and semantic analysis.

Putnam, H. 1967. The nature of mental states. In (Capitan & Merrill, eds) *Art, Mind, and Religion*. Pittsburgh University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Why mental states are more likely to be functional states (in probabilistic automata) than brain states or behavioral dispositions.

Putnam, H. 1967. The mental life of some machines. In (H. Castaneda, ed) *Intentionality, Minds and Perception*. Wayne State University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

On explaining behavior via TM states, e.g. explaining preference via utility functions. Logical behaviorism assumes rational preference functions. Functional organization is what matters, not physical make-up.

Putnam, H. 1975. Philosophy and our mental life. In *Mind, Language, and Reality*. Cambridge University Press.

Psychological states aren't TM states after all: we have lots of psych states at once; they depend on learning/memory; disjunctions of TM states are no good. But functional organization rather than physics is still what counts.

Putnam, H. 1987. *Representation and Reality*. MIT Press.

Type functionalism isn't any better than type physicalism, as mental states can be multiply realized as functional states. With what in common?

Lycan, W.G. 1974. Mental states and Putnam's functionalist hypothesis. *Australasian Journal of Philosophy* 52:48-62.

On abstract vs. physical TMs: Putnam should say that mental states are physical TM states. But then functionalism is compatible with physicalism. On the relation between Putnam's and Armstrong's functionalism.

Lycan, W.G. 1979. A New Lilliputian argument against machine functionalism. *Philosophical Studies* 35:279-87.

If machine functionalism were true, a homunculus-head would have all the mental states of its homunculus (by the definition of "realization"), which is absurd.

Lycan, W.G. 1983. The moral of the New Lilliputian argument. *Philosophical Studies* 43:277-80.

Reply to Elugardo 1983: so how do you specify what count as inputs/outputs?

Elugardo, R. 1981. Machine functionalism and the New Lilliputian argument. *Pacific Philosophical Quarterly* 62:256-61.

Criticism of Lycan 1979, and a re-making of the argument.

Elugardo, R. 1983. Machine realization and the New Lilliputian argument. *Philosophical Studies* 43:267-75.

Lycan's New Lilliputian argument fails as inputs/outputs for the homunculus are not the same as inputs/outputs for the full system.

Kane, R.H. 1966. Turing machines and mental objects

Nelson, R. 1974. Mechanism, functionalism, and the identity theory. *Journal of Philosophy* 73:365-86.

Argues for mechanism rather than functionalism. Criticizes Putnam for hypostasizing mental states, which are disanalogous to mental states. Defending mechanism against Kalke's & Rorty's objections.

Rorty, R. 1972. Functionalism, machines and incorrigibility. *Journal of Philosophy* 69:203-20.

Logical states don't give us any understanding of mind over and above what the function/structure distinction gives us. In particular, it doesn't help with the understanding of privacy and incorrigibility.

Tomberlin, J. 1965. About the identity theory. *Australasian Journal of Philosophy* 43:295-99.

Contra Putnam: logical states are not physical states, and utterances about them are not about physical states.

Wagner, S.J. 1988. The liberal and the lycanthrope. *Pacific Philosophical Quarterly* 69:165-74.

Contra Lycan: machine functionalism can handle Bolivia and CRT cases by a causal/counterfactual account, and Lilliputian case by assigning mental states to minds, not bodies.

3.4c Functionalism, Miscellaneous

Adams, F. 1979. Properties, functionalism, and the identity theory. *Eidos* 1:153-79.

Bealer, G. 1978. An inconsistency in functionalism. *Synthese*.

A formal argument showing that functional definitions are equivalent to behavioral definitions.

Bealer, G. 1985. Mind and anti-mind: Why thinking has no functional definition. *Midwest Studies in Philosophy* 9:283-328.

Bechtel, W. 1984. Autonomous psychology: What it should and should not entail. *Philosophy of Science Association* 1984, 1:43-55.

The functional level is the appropriate level for psychology, but neurophysiological facts constrain this level and are thus relevant.

Ben-Yami, H. 1999. An argument against functionalism. *Australasian Journal of Philosophy* 77:320-324.

Biro, J.I. & Shahan, R.W. (eds) 1982. *Mind, Brain and Function*. Oklahoma University Press.

Ten papers on functionalism. Originally was *Philosophical Topics*, volume 12.

Block, N. 1980. Functionalism. In (N. Block, ed) *Readings in the Philosophy of Psychology*, Vol. 1. MIT Press.

Distinguishes varieties of functionalism, e.g. machine and Ramsey-sentence functionalism; and compares to behaviorism. With a historical overview, and arguments for why functionalism is incompatible with physicalism.

Block, N. 1978. Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9:261-325. Reprinted in *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Distinguishes analytic and empirical functionalism. Both have problems with absent qualia, and inputs/outputs. Analytic functionalism has problems with paralytics, etc; empirical functionalism has problems with Martians.

Block, N. & Fodor, J.A. 1972. What psychological states are not. *Philosophical Review* 81:159-81. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

Mental states are not physical or behavioral states; could they be functional states? With various arguments against type identity, and against machine-table functionalism.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72:741-64. Reprinted in (N. Block, ed) *Readings in the Philosophy of Psychology* (MIT Press, 1980).

On the role of functional explanation versus other kinds of explanation. Functionalism applies an analytic, not subsumptive strategy.

David, M. 1997. Kim's functionalism. *Philosophical Perspectives* 11:133-48.

Fischer, J. 1985. Functionalism and propositions. *Philosophical Studies* 48:295-311.

Fodor, J.A. 1968. Materialism. In *Psychological Explanation*. Random House.

On mental state as inferred theoretical entities, individuated according to their function (cf. valve-lifters). Psychology and neuroscience will mutually constrain each other, giving a relation more complex than reduction.

Gendron, B. 1970. On the relation of neurological and psychological theories: A critique of the hardware thesis. *Boston Studies in the Philosophy of Science* 8:483-95.

Argues that functional explanation are reducible to structural explanations.

Hornsby, J. 1986. Physicalist thinking and conceptions of behaviour. In (P. Pettit & J. McDowell, eds) *Subject, Thought, and Context*. Oxford University Press.

Hoy, R.C. 1980. Dispositions, logical states, and mental occurrents. *Synthese* 44:207-40.

Kalke, W. 1969. What's wrong with Fodor's and Putnam's functionalism. *Nous* 3:83-93.

There's no absolute functional/structural distinction, as it depends on how you choose boundaries and levels of abstraction.

Lycan, W.G. 1981. Form, function and feel. *Journal of Philosophy* 78:24-50.
Pursue a multi-leveled homuncular functionalism, with mental states characterized as states of teleologically identified subsystems. Even the identity theorist is a functionalist at a low level.

Malcolm, N. 1980. 'Functionalism' in philosophical psychology. *Proceedings of the Aristotelian Society* 80:211-30.

Pereboom, D. 1991. Why a scientific realist cannot be a functionalist. *Synthese* 88:341-58.

Scientific realism requires dispositions of kinds be explained by intrinsic properties. Neural/functional properties won't work, because of reductionism and circularity. Use intrinsic psychological properties instead.

Richardson, R.C. 1979. Functionalism and reductionism. *Philosophy of Science* 46:533-58.

Argues that functionalism is compatible with reductionism, by analogies. Genetics has multiple realization and multiple function; reduction doesn't require biconditionals. With remarks on the de facto autonomy of psychology.

Schiffer, S. 1986. Functionalism and belief. In (M. Brand & R. Harnish, eds) *The Representation of Knowledge and Belief*. University of Arizona Press.

Against functionalism for beliefs. Both common-sense functionalism and psychofunctionalism have problems with finding the right functional theory, distinguishing beliefs, perceptual input conditions, Twin Earth, etc.

Shope, R.K. 1973. Functional equivalence and the defense of materialism. *Philosophical Forum* 4:500-12.

Sober, E. 1990. Putting the function back into functionalism. In (W. Lycan, ed) *Mind and Cognition*. Blackwell.

Need teleological functionalism, not Turing Machine functionalism.

Sober, E. 1985. Panglossian functionalism and the philosophy of mind. *Synthese* 64:165-93.

van Gulick, R. 1982. Functionalism as a theory of mind. *Philosophy Research Archives* 185-204.

The structure/function distinction is level-relative, so physiology might be relevant even under functionalism. Problems with automata, and with causal connections to nonintentionally characterized behavior.

van Gulick, R. 1980. Functionalism, information and content. *Nature and System* 2:139-62. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Ward, A. 1989. Philosophical functionalism. *Behaviorism* 17:155-8.

Weckert, J. 1990. Functionalism's impotence. *Philosophical Inquiry* 32-43.

Wilkes, K.V. 1981. Functionalism, psychology and the philosophy of mind. *Philosophical Topics* 12:147-67.

Functionalism may be appropriate for cognitive psychology but not for folk psychology, due to differing goals. Neuroscience will play an important role in developing functional theories.

Zangwill, N. 1992. Variable realization: not proven. *Philosophical Quarterly* 42:214-19.

Argues that the possibility of multiple realization has not been established, whether by arguments from imagination, concepts, or empirical facts.

3.5 Other Psychophysical Theories

3.5a Logical Behaviorism (Ryle, etc)

- Ryle, G. 1949. *The Concept of Mind*. Hutchinson and Co.
The ancestor of most contemporary philosophy of mind. Among other things, argues that the "ghost in the machine" view of mind is a category mistake, and presents dispositional analyses of many mental concepts.
- Bestor, T.W. 1979. Gilbert Ryle and the adverbial theory of mind. *Personalist* 60:233-42.
- Campbell, C.A. 1953. Ryle on the intellect. *Philosophical Quarterly* 3:115-38.
- Carnap, R. 1959. Psychology in physical language. In (Ayer, ed) *Logical Positivism*. Free Press.
- Carrier, L. 1973. Professor Shaffer's refutation of behaviourism. *Mind* 80:249-52.
- Chisholm, R. 1955. A note on Carnap's meaning analysis. *Philosophical Studies*.
- Chisholm, R. 1952. Intentionality and the theory of signs. *Philosophical Studies*.
- Chisholm, R. 1958. Sentences about believing. *Minnesota Studies in the Philosophy of Science* 2.
- Dalrymple, H. 1977. Some logical muddles in behaviorism. *Southwestern Philosophical Studies* 2:64-72.
- Ewing, A.C. 1953. Professor Ryle's attack on dualism. *Proceedings of the Aristotelian Society* 53:47-78.
- Farrell, B. 1950. Experience. *Mind* 59:170-98.
- Finn, D.R. 1971. Putnam and logical behaviourism. *Mind* 80:432-36.
- Flanagan, O.J. & McCreadie-Albright, T. 1974. Malcolm and the fallacy of behaviorism. *Philosophical Studies* 26:425-30.
- Geach, P. 1957. *Mental Acts*. Routledge and Kegan Paul.
- Goudge, T.A. 1982. Ryle's last thoughts on thinking. *Dialogue* 21:125-32.
- Graham, G. 19xx. Spartans and behaviorists. *Behaviorism* x:xx.
Defends behaviorism as a scientific hypothesis, so that conceivability arguments aren't relevant, and advocates "penetrability" behaviorism which can appeal to internal physical states.
- Hamer, C. 1970. Why Ryle is not a behaviourist. *Philosophical Studies (Ireland)* 17:7-25.
- Hamlyn, D.W. 1953. Behaviour. *Philosophy* 28:132-45.
- Hanson, N.R. 1952. Professor Ryle's "mind". *Philosophical Quarterly* 2:246-48.
- Heidelberger, H. 1966. On characterizing the psychological. *Philosophy and Phenomenological Research*.
- Kitchener, R.F. 1977. Behavior and behaviorism. *Behaviorism* 5:11-68.
- Jacquette, D. 1985. Logical behaviorism and the simulation of mental episodes.

Journal of Mind and Behavior 6:325-332.

Mace, C.A. 1949. Some implications of analytical behaviourism. Proceedings of the Aristotelian Society.

Malcolm, N. 1954. Wittgenstein's *Philosophical Investigations*. Philosophical Review 43:530-9.

Mandelbaum, M. 1958. Professor Ryle and psychology. Philosophical Review 67:522-30.

McLaughlin, B. & O'Leary-Hawthorne, J. 1995. Dennett's logical behaviorism. Philosophical Topics 22:189-258.

Miller, D.S. 1911. Is consciousness "a type of behaviour"? Journal of Philosophy 8:322-27.

Miller, D.S. 1951. "Descartes myth" and "Professor Ryle's fallacy". Journal of Philosophy.

Nelson, R. 1969. Behaviorism is false. Journal of Philosophy 66:417-52.

Nelson, R. 1975. Behaviorism, finite automata, and stimulus-response theory. Theory and Decision 6:249-67.

Oosthuizen, D.C.S. 1970. Phenomenological psychology. Mind 79:487-501.

Park, S. 1994. Reinterpreting Ryle: A nonbehaviorist analysis. Journal of the History of Philosophy 32:265-90.

Place, U.T. 1993. A radical behaviorist methodology for the empirical investigation of private events. Behavior and Philosophy 20:25-35.

Price, H.H. 1960. Some objections to behaviorism. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

Putnam, H. 1963. Brains and behavior. In (R. Butler, ed) *Analytical Philosophy: Second Series*. Blackwell. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Quine, W.V. 1975. Mind and verbal dispositions. In (Guttenplan, ed) *Mind and Language*. Oxford University Press.

Quine, W.V. 1980. Sellars on behaviorism, language, and meaning. Pacific Philosophical Quarterly 61:26-30.

Robinson, H. 1982. Behaviorism and stimulus materialism. In *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Rowlands, M. 1991. A defense of behaviorism. Behavior and Philosophy 19:93-100.

Ryle, G. 1979. *On Thinking*. Blackwell.

Scriven, M. 1956. A study of radical behaviorism. Minnesota Studies in the Philosophy of Science 1:88-130.

Sellars, W. 1952. Mind, meaning, and behavior. Philosophical Studies.

Shuford, H. 1966. Logical behaviorism and intentionality. Theoria 32:246-51.

Skinner, B.F. 1945. The operational analysis of psychological terms. Psychological Review 52:270-78.

Smart, J.J.C. 1959. Ryle on mechanism and psychology. *Philosophical Quarterly* 9:349-55.

Stemmer, N. 1993. Behavioral materialism, the success of folk psychology, and the first-person case. *Behavior and Philosophy* 20:1-14.

Vendler, Z. 1981. Ryle's thoughts on thinking. *Midwest Studies of Philosophy* 6:335-43.

Weitz, M. 1951. Professor Ryle's "logical behaviourism". *Journal of Philosophy* 48:297-300.

Whitely, C.A. 1961. Behaviourism. *Mind* 70:164-74.

Wisdom, J. 1950. The concept of mind. *Proceedings of the Aristotelian Society* 50:189-204.

Wittgenstein, L. 1953. *Philosophical Investigations*.

Wright, J.N. 1959. Mind and the concept of mind. *Aristotelian Society Supplement* 33:1-22.

Ziff, P. 1958. About behaviourism. *Analysis* 18:132-6.

3.5b Identity Theory (Smart, etc) [see also 1.3c, 1.4g, 3.3b]

Feigl, H. 1958. The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science* 2:370-497. Reprinted as *The 'Mental' and the 'Physical'*. University of Minnesota Press, 1967.

Place, U.T. 1956. Is consciousness a brain process? *British Journal of Psychology* 47:44-50. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

The idea that consciousness is a brain process is logically coherent. It's a scientific hypothesis, not a necessary truth. On the "is" of composition vs the "is" of definition, and the fallacy of the internal phenomenal field.

Smart, J.J.C. 1959. Sensations and brain processes. *Philosophical Review* 68:141-56.

Defending the thesis that sensations are contingently identical to brain processes against various objections. Topic-neutral analysis of sensation reports. Materialism beats epiphenomenalism on grounds of simplicity.

Abelson, R. 1970. A refutation of mind-body identity. *Philosophical Studies* 18:85-90.

The number of possible mental states is infinite (think of any number), whereas there are only finitely many brain states, so they're not identical.

Armstrong, D.M. 1968. The headless woman and the defense of materialism. *Analysis* 29:48-49.

Likens the anti-materialist position to the "headless woman" fallacy: "I'm not aware the mental states are physical", so "I'm aware that mental states are non-physical".

Armstrong, D.M. 1973. Epistemological foundations for a materialist theory of mind. *Philosophy of Science* 40:178-93.

A prima facie case for materialism based on grounds of rational consensus, arising especially from common-sense and scientific evidence. Mental states exist (common-sense) but should be analyzed causally (evidence from science).

Aune, B. 1966. Feigl on the mind-body problem. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in*

Honor of Herbert Feigl_. University of Minnesota Press.

Baier, K. 1962. Smart on sensations. Australasian Journal of Philosophy 40:57-68.

Mental states are necessarily private, and so cannot be physical states, which are public. We have epistemological authority about our mental states.

Beloff, J. 1965. The identity hypothesis: A critique. In (J.R. Smythies, ed) _Brain and Mind_. Routledge and Kegan Paul.

Blumenfeld, J-B. 1979. Phenomenal properties and the identity theory. Australasian Journal of Philosophy 63:485-93.

Argues that phenomenal properties aren't needed to identify sensations with brain-states, and nor are topic-neutral analyses.

Borst, C.V. (ed) 1970. _The Mind/Brain Identity Theory_. Macmillan.
An anthology of central articles on the identity theory.

Bradley, M.C. 1963. Sensations, brain-processes, and colours. Australasian Journal of Philosophy 41:385-93.

Brandt, R. 1960. Doubts about the identity theory. In (S. Hook, ed) _Dimensions of Mind_. New York University Press.

Brandt, R. & Kim, J. 1967. The logic of the identity theory. Journal of Philosophy 66:515-537.

Arguing for an event-identity construal of the identity theory. Comparing the identity theory to the weaker "principle of simultaneous isomorphism". The only reason to accept the identity theory is ontological simplicity.

Brodbeck, M. 1966. Mental and physical: Identity versus sameness. In (P. Feyerabend & G. Maxwell, eds) _Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl_. University of Minnesota Press.

Candlish, S. 1970. Mind, brain, and identity. Mind 79:502-18.

Carney, J. 1971. The compatibility of mind-body identity with dualism. Mind. Argues that the identity theory is compatible with linguistic dualism, as the mental and the physical may differ in intensional properties only.

Clarke, J. 1971. Mental structure and the identity theory. Mind 80:521-30.

Coburn, R. 1963. Shaffer on the identity of mental states and brain processes. Journal of Philosophy 60:89-92.

Coder, D. 1973. The fundamental error of central-state materialism. American Philosophical Quarterly 10:289-98.

On problems with theories that leave the nature of mind open a priori: how can we even understand the possibilities?

Cornman, J. 1962. The identity of mind and body. Journal of Philosophy 59:486-92.

Coburn, R.C. 1963. Shaffer on the identity of mental states and brain processes. Journal of Philosophy 60:89.

Location of mental states by convention (Shaffer 1961) won't work, as it (a) makes mental states public, and (b) conflicts with connections to behavior.

Crittenden, C. 1971. Ontology and mind-body identity. Philosophical Forum 2:251-70.

de Boer, R. 1976. Cartesian categories in mind-body identity theories.

Philosophical Forum 7:139-58.

Double, R. 1981. Central state materialism. Philosophical Studies (Ireland) 28:229-37.

Enc, B. 1983. In defense of the identity theory. Journal of Philosophy 80:279-98.

Feigl, H. 1971. Some crucial issues of mind-body monism. Synthese.

Garnett, A.C. 1965. Body and mind: the identity thesis. Australasian Journal of Philosophy 43:77-81.

Grunbaum, A. 1972. Abelson on Feigl's mind-body identity thesis. Philosophical Studies 23:119-21.

Gustafson, D.F. 1963. On the identity theory. Analysis 24:30-32.

Hanratty, G. 1972. The identity theory of Herbert Feigl. Philosophical Studies 20:113-23.

Harris, E.E. 1966. The neural identity thesis and the person. International Philosophical Quarterly 6:515-37.

Hedman, C.G. 1970. On correlating brain states with psychological states. Australasian Journal of Philosophy 48:247-51.

Heil, J. 1970. Sensations, experiences, and brain processes. Philosophy 45:221-6.

Hinton, J.M. 1967. Illusions and identity. Analysis 27:65-76.

Hockutt, M. 1967. In defense of materialism. Philosophy and Phenomenological Research 27:366-85.

Hoffman, R. 1967. Malcolm and Smart on brain-mind identity. Philosophy 42:128-36.

Joske, W. 1960. Sensations and brain processes: A reply to Professor Smart. Australasian Journal of Philosophy 38:157-60.

On topic-neutral reports, after-images, and after-radishes. Such a report requires epistemic access to physical resemblance, which we don't have.

Kim, J. 1966. On the psycho-physical identity theory. American Philosophical Quarterly 3:227-35.

There's no empirical support for identity, over and above that for correlation; and unity of science gives no reason to accept identity. The only reason might be that of ontological simplicity.

Kim, J. 1972. Phenomenal properties, psychophysical laws and the identity theory. Monist 56:178-92.

Deal with phenomenal properties by allowing only mental events, and eliminating mental objects. Identity theories needn't suppose psychophysical laws. With defense against multiple realizability arguments.

Kitcher, P.S. 1982. Two versions of the identity theory. Erkenntnis 17:213-28.

Recasting the identity theory and functionalism, using Kripkean theories of reference, so mental states can refer to physiological or psychological states that we don't yet understand; and qualia problems are handled better.

Lewis, D. 1965. An argument for the identity theory. Journal of Philosophy 63:17-25. Reprinted in *Philosophical Papers, Vol. 1* (Oxford University Press, 1980).

Mental states are defined by their causal roles. So, by the completeness of physics, they must be physical states.

Locke, D. 1971. Must a materialist pretend he's anaesthetized? *Philosophical Quarterly* 49:217-31.

On how materialism, as opposed to a double aspect view, can handle mental features -- by moving them into the world via a realist theory of perception. Remarks on identification of states. After-images, etc, cause problems.

Lockwood, M. 1984. Einstein and the identity theory. *Analysis*.

Using the special theory of relativity to show that if mental events have a temporal location, then they must have a spatial location.

Lubow, N. 1978. Mind-body identity and irreducible properties. *Philosophy Research Archives* 4:1240.

Luce, D.R. 1966. Mind-body identity and psycho-physical correlation. *Philosophy of Science* 17:1-7.

Malcolm, N. 1964. Scientific materialism and the identity theory. *Dialogue* 3:115-25.

The identity theory is meaningless, if identity is analyzed as spatiotemporal coincidence, as thoughts don't have location. Thoughts also require context. Even if identity holds, explaining brain doesn't imply explaining mind.

Macdonald, C. 1989. *Mind-Body Identity Theories*. Routledge.

Malcolm, N. 1964. Scientific materialism and the identity theory. *Dialogue*.

Margolis, J. 1965. Brain processes and sensations. *Theoria* 31:133-38.

Meehl, P. 1966. The compleat autocerebroscopist: A thought-experiment on Professor Feigl's mind-body identity thesis. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Mucciolo, L. 1974. The identity theory and criteria for the mental. *Philosophy and Phenomenological Research* 35:167-80.

Munsat, S. 1969. Could sensations be processes? *Mind* 78:247-51.

Sensations and processes have different logical type, so it is a priori impossible that they should be identical.

Nagel, T. 1965. Physicalism. *Philosophical Review* 74:339-56, 1965.

Noren, S.J. 1970. Identity, materialism, and the problem of the danglers. *Metaphilosophy* 4:318-44.

Noren, S.J. 1970. Smart's materialism: The identity thesis and translation. *Australasian Journal of Philosophy* 48:54-66.

Norton, R. 1964. On the identity of identity theories. *Analysis* 25:14-16.

Pepper, S. 1975. A split in the identity theory. In (C. Cheng, ed) *Philosophical Aspects of the Mind-Body Problem*. Hawaii University Press.

Pitcher, G. 1960. Sensations and brain processes: A reply to Professor Smart. *Australasian Journal of Philosophy* 38:150-7.

Identity requires explanation to be accepted, but Smart doesn't provide this. But one can deny identity without claiming dualism -- e.g. a "duck-rabbit" theory of mind/brain. With remarks on the completeness of descriptions.

Place, U.T. 1960. Materialism as a scientific hypothesis. *Philosophical*

Review 69:101-4.

Contra Smart 1959: Materialism is a scientific hypothesis, if we accept certain logical criteria for what a sensation is; otherwise it's just false.

Place, U.T. 1972. Sensations and processes: A reply to Munsat. *Mind*.

Place, U.T. 1988. Thirty years on -- Is consciousness still a brain process? *Australasian Journal of Philosophy* 66:208-19.

Comparing contemporary materialism to Pace's 1956 variety. With remarks on whether the thesis is empirical or a priori, and on deciding the issue between materialism and epiphenomenalism.

Place, U.T. 1989. Low claim assertions. In (J. Heil, ed) *Cause, Mind, and Reality: Essays Honoring C. B. Martin*. Kluwer.

Discusses a paper of Martin's and the genesis of the identity theory, with a focus on 'public' and 'private logic' and topic-neutral descriptions.

Presley, C.P. (ed) 1967. *The Identity Theory of Mind*. University of Queensland Press.

Puccetti, R. 1978. The refutation of materialism. *Canadian Journal of Philosophy* 8:157-62.

The identity theory must be false, as pain centers in vitro will not be pains. With a reply by G. Pearce and a rejoinder.

Ripley, C. 1969. The identity theory and scientific hypotheses. *Dialogue* 2:308-10.

Robinson, H. 1982. The disappearance theory. In *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge University Press.

Rosenbaum, S. 1977. The property objection and the principles of identity. *Philosophical Studies* 32.

Routley, R. & McCrae, V. 1966. On the identity of sensations and physiological occurrences. *American Philosophical Quarterly* 3.

Schlagel, R.H. 1977. The mind-body identity impasse. *American Philosophical Quarterly* 14:231-37.

Scriven, M. 1966. The limitations of the identity theory. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

On the identity theory as a linguistic proposal, compatible with dualism; epiphenomenalism and parallelism must be false, leaving interactionism.

Sellars, W. 1965. The identity approach to the mind-body problem. *Review of Metaphysics* 18:430-51.

Shaffer, J. 1961. Could mental states be brain processes? *Journal of Philosophy* 58:813-22.

Mental states don't have a location, and brain processes do; but we could stipulate a location for mental states. With remarks on possible relations between mental and physical features, states, and concepts.

Shaffer, J. 1963. Mental events and the brain. *Journal of Philosophy* 60:160-6.

We identify mental events by noticing mental features that must be nonphysical, but still might be empirically reducible. Against topic-neutral definitions, and with response to Coburn 1963 on location.

Simon, M.A. 1970. Materialism, mental language, and the mind-body identity. *Philosophy and Phenomenological Research* 30:514-32.

Smart, J.J.C. 1960. Sensations and brain processes: A rejoinder to Dr. Pitcher and Mr. Joske. *Australasian Journal of Philosophy* 38:252-54.

Smart, J.J.C. 1961. Further remarks on sensations and brain processes. *Philosophical Review*.

Reply to Stevenson 1960: There are no irreducible mental properties; they reduce to physical properties via topic-neutral definitions.

Smart, J.J.C. 1962. Brain processes and incorrigibility. *Australasian Journal of Philosophy* 40:68-70.

Reply to Baier 1962: epistemological authority is compatible with materialism. Mental state reports are not completely incorrigible, though.

Smart, J.J.C. 1963. Materialism. *Journal of Philosophy* 60:651-62.

Defending topic-neutral analyses of mental reports, and arguing against Wittgensteinian behaviorism via brain-in-vat examples. With remarks on the appeal of materialism and on compatibility with ordinary language.

Smart, J.J.C. 1965. The identity thesis: A reply to Professor Garrett. *Australasian Journal of Philosophy* 43:82-3.

Smart, J.J.C. 1972. Further thoughts on the identity theory. *Monist* 56:177-92.

On some problems for the identity theory arising from the intensionality of mental states and from the appeal to properties, and on how to modify the translation form of the theory without embracing the disappearance version.

Smythies, J.R. 1994. Requiem for the identity theory. *Inquiry* 37:311-29.

Sosa, E. 1965. Professor Malcolm on "Scientific materialism and the identity theory". *Dialogue* 4:422-23.

Stevenson, J.T. 1960. 'Sensations and brain processes': A reply to J.J.C. Smart. *Philosophical Review* 69:505-10.

Identity theory implies nomological danglers, due to the irreducibility of defining mental properties.

Stoutland, F. 1971. Ontological simplicity and the identity hypothesis. *Philosophy and Phenomenological Research*.

The identity thesis isn't ontologically simpler than dualism: we still need a dualism of properties, and explanatory danglers. Not much turns on the issue, except in teleological explanation.

Sosa, E. 1965. Professor Malcolm on 'Scientific materialism and the identity theory'. *Dialogue* 3:422-23.

Contra Malcolm 1965: explaining brain will explain mind, if the explanation is conjoined with the identity statement. With rejoinder from Malcolm.

Swartz, N. 1974. Can the theory of contingent identity between sensation-states and brain-states be made empirical? *Canadian Journal of Philosophy* 3:405-17.

Swinburne, R. 1993. Are mental events identical with brain events? *American Philosophical Quarterly* 19:173-181.

Property identity theses fail due to meaning differences, and event identity these fail due to a lack of entailment relations. Rebuts objections from weaker identity criteria and analogies with scientific identification.

Taylor, C. 1967. Mind-body identity, a side issue? *Philosophical Review* 76:201-13.

Teichmann, J. 1967. The contingent identity of minds and brains. *Mind*

76:404-15.

Thalberg, I. 1978. A novel approach to mind-brain identity. *Philosophy of Science* 3:255-72.

Suggests a theory in which neural states are components of, but not identical to, overall psychological states. This can accommodate raw feels if necessary as a further component, but is mostly materialistic.

Thomson, J.J. 1969. The identity theory. In (S. Morgenbesser, P. Suppes, & M. White, eds) *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*. St. Martin's Press.

Tomberlin, J.E. 1965. About the identity theory. *Australasian Journal of Philosophy* 53:295-9.

Watkins, J.W.N. 1978. A basic difficulty in the mind-brain identity hypothesis. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Weismann, D. 1965. A note on the identity thesis. *Mind* 74:571-77.

Whitely, C.H. 1970. The mind-brain identity hypothesis. *Philosophical Quarterly* 20:193-99.

Wolfe, J. & Nathan, G.J. 1968. The identity theory as a scientific hypothesis. *Dialogue* 7:469-72.

Ziedins, R. 1971. Identification of characteristics of mental events with characteristics of brain events. *American Philosophical Quarterly* 8:13-23.

3.5c Eliminative Materialism (Rorty, Feyerabend) [see also 1.4d, 1.7c, 2.1c]

Austin, J.W. 1975. Rorty's materialism. *Auslegung* 3:20-28.

Bernstein, R. 1968. The challenge of scientific materialism. *International Philosophical Quarterly* 8:252-75.

Bush, E. 1974. Rorty revisited. *Philosophical Studies* 25:33-42.

Cam, P. 1978. "Rorty revisited", or "Rorty revised". *Philosophical Studies* 33:377-86.

Carter, W.R. 1974. On incorrigibility and eliminative materialism. *Philosophical Studies* 28:113-21.

Cornman, J. 1968. On the elimination of 'sensations' and sensations. *Review of Metaphysics* 22:15-35.

Donovan, C. 1978. Eliminative materialism reconsidered. *Canadian Journal of Philosophy* 8.

Doppelt, G. 1977. Incorrigibility, the mental, and materialism. *Philosophy Research Archives*.

Everitt, N. 1981. A problem for the eliminative materialist. *Mind* 90:428-34.

Everitt, N. 1983. How not to solve a problem for the eliminative materialist. *Mind* 92:590-92.

Feyerabend, P. 1963. Mental events and the brain. *Journal of Philosophy* 40:295-6. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990). Identity theory implies dualism, though its acceptance of mental properties. Instead we should eliminate talk of mental processes altogether, or redefine them in physiological terms.

Feyerabend, P. 1963. Materialism and the mind-body problem. Review of Metaphysics 17:49-67.

Globus, G. 1989. The strict identity theory of Schlick, Russell, Maxwell, and Feigl. In (M. Maxwell & C. Savage, eds) *Science, Mind, and Psychology: Essays in Honor of Grover Maxwell*. University Press of America.

Godow, R. 1976. Eliminative materialism and denotation. Philosophy and Phenomenological Research 36.

Goodman, R.B. 1974. A note on eliminative materialism. Journal of Critical Analysis 5:80-83.

Hiley, D.R. 1978. Is eliminative materialism materialistic? Philosophy and Phenomenological Research 38:325-37.

Hiley, D.R. 1980. The disappearance theory and the denotation argument. Philosophical Studies 37:307-20.

Lycan, W.G. & Pappas, G. 1972. What is eliminative materialism? Australasian Journal of Philosophy 50:149-59.

Lycan, W.G. 1976. Quine's materialism. Philosophia 6:101-30.

Quine, W.V. 1966. On mental entities. In *The Ways of Paradox*. Random House.

Richardson, R.C. 1981. Disappearance and the identity theory. Canadian Journal of Philosophy 11:473-85.

Rorty, R. 1965. Mind-body identity, privacy, and categories. Review of Metaphysics 19:24-54.

Rorty, R. 1970. Incorrigibility as the mark of the mental. Journal of Philosophy.

Rorty, R. 1970. In defense of eliminative materialism. Review of Metaphysics 24:112-21.

Rosenthal, D.M. 1980. Keeping matter in mind. Midwest Studies in Philosophy 5:295-322.

Savitt, S. 1974. Rorty's disappearance theory. Philosophical Studies 28:433-36.

Shirley, E.S. 1974. Rorty's "disappearance" version of the identity theory. Philosophical Studies 25:73-75.

Sikora, R.I. 1974. Rorty's mark of the mental and his disappearance theory. Canadian Journal of Philosophy 4:191-93.

Sikora, R.I. 1975. Rorty's new mark of the mental. Analysis 35:192-94.

Steiling, K. 1976. The elimination of sensations and the loss of philosophy. Auslegung 3:20-28.

3.5d Anomalous Monism (Davidson) -----

Davidson, D. 1970. Mental events. In (L. Foster & J. Swanson, eds) *Experience and Theory*. Humanities Press. Reprinted in *Essays on Action and Events* (Oxford University Press, 1980).

Arguing for anomalous monism: no strict psychophysical laws, no strict

psychological laws, and token identity without type identity. Mental events can still cause, via subsumption under physical laws.

Davidson, D. 1973. The material mind. In (P. Suppes, ed) *Logic, Methodology and the Philosophy of Science*. North-Holland. Reprinted in *Essays on Action and Events* (Oxford University Press, 1980).

The psychological supervenes on the physical but is not reducible to it, because of the holistic nature of intentional attribution. So building a perfect physical model may not explain psychology.

Davidson, D. 1974. Psychology as philosophy. In (S. Brown, ed) *Philosophy of Psychology*. Harper & Row. Reprinted in *Essays on Action and Events* (Oxford University Press, 1980).

On the differing constitutive standards of mental and physical concepts. Attribution of mental concepts is holistic, and presupposes a background of rationality, etc. With examples from decision theory.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford University Press. A collection of papers on action, causation and the philosophy of psychology.

Davidson, D. 1987. Problems in the explanation of action. In (P. Pettit, R. Sylvan, & J. Norman, eds) *Metaphysics and Morality*. Blackwell. Remarks on how mental properties can explain action without strict laws. The mental is a conceptual, not an ontological category, governed by normative standards, and not reducible to the non-normative.

Davidson, D. 1992. Thinking causes. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Davidson, D. 1995. Laws and cause. *Dialectica* 49:263-79.

Antony, L. 1989. Anomalous monism and the problem of explanatory force. *Philosophical Review* 98:153-87.

Criticism of Davidson's argument for rational causation. Reasons must cause in virtue of their rational properties. Token identities can't exist, due to normativity. Quinean psychology can't yield rational explanations.

Bickle, J. 1992. Mental anomaly and the new mind-brain reductionism. *Philosophy of Science* 59:217-30.

Campbell, N. 1997. The standard objection to anomalous monism. *Australasian Journal of Philosophy* 75:373-82.

Campbell, N. 1998. Anomalous monism and the charge of epiphenomenalism. *Dialectica* 52:23-39.

Cheng, K. 1997. Davidson's action theory and epiphenomenalism. *Journal of Philosophical Research* 22:81-95.

Child, W. 1993. Anomalism, uncodifiability, and psychophysical relations. *Philosophical Review*.

Anomalism is compatible with supervenience, if it is construed as denying psychophysical laws useful for explaining behavior. It is incompatible with token identity, though. With much on the uncodifiability of rationality.

Cooper, W.E. 1980. Materialism and madness. *Philosophical Papers* 9:36-40.

Daniel, S.G. 1999. Why even Kim-style psychophysical laws are impossible. *Pacific Philosophical Quarterly* 80:225-237.

Elgin, C. 1980. Indeterminacy, underdetermination and the anomalous monism. *Synthese* 45:233-55.

Garrett, B. 1999. Davidson on causal relevance. *Ratio* 12:14-33.

Goldberg, B. 1977. A problem with anomalous monism. *Philosophical Studies* 32:175-80.

Davidson's argument equivocates on the term "physical": the physical events that mental events cause might not be subsumed under laws.

Hess, P. 1981. Actions, reasons and Humean causes. *Analysis* 41:77-81.

Anomalous monism implies that mental properties don't cause anything.

Honderich, T. 1982. The argument for anomalous monism. *Analysis* 42:59-64.

If anomalous monism is true, mental events may cause, but their mental properties aren't causally relevant.

Johnston, M. 1985. Why having a mind matters. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

Anomalous monism loses out to Australian materialism. It can't be a priori, it leads to exhaustive monism, it doesn't support a new view of free action, and it implies the causal irrelevance of the mental.

Kalderon, M.E. 1987. Epiphenomenalism and content. *Philosophical Studies* 52:71-90.

Davidson's view leads to epiphenomenalism about content, as it can't support the appropriate counterfactuals. Strong supervenience might be a way out, but that is inconsistent with anomalism.

Kernohan, A. 1985. Psychology: Autonomous or anomalous? *Dialogue* 24:427-42.

Kim, J. 1985. Psychophysical laws. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).

How there can be psychophysical generalizations but no laws -- they might lack modal force. On the relation between psychophysical anomalism and psychological anomalism. Casting Davidson as a Kantian dualist.

Kim, J. 1993. Can supervenience and "non-strict laws" save anomalous monism? In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Klagge, J.C. 1990. Davidson's troubles with supervenience. *Synthese* 85:339-52.

Anomalous supervenience is consistent, at the cost of anti-realism about the mental. Supervenience is a constraint on interpretation, but needn't support counterfactuals as different interpretation schemes are possible,

Klee, R. 1992. Anomalous monism, ceteris paribus, and psychological explanation. *British Journal for the Philosophy of Science* 43:389-403.

Problems with holism and ceteris paribus laws aren't unique to psychology. One finds the same thing in the physical sciences. So rationality plays no special role, and psychological laws are as reasonable as physical laws.

Kuczynski, J.M. 1998. A proof of the partial anomalousness of the mental. *Southern Journal Of Philosophy* 36:491-504.

Latham, N. 1999. Davidson and Kim on psychophysical laws. *Synthese* 118:121-44.

LePore, E. & Loewer, B. 1987. Mind matters. *Journal of Philosophy* 630-42.

Anomalous monism is not committed to epiphenomenalism, as even non-strict laws can ground counterfactuals and so support the causal relevance of mental properties.

Lycan, W.G. 1981. Psychological laws. *Philosophical Topics* 12:9-38.

A functionalist defense against anomalous monism. Psychofunctional laws

and psychological laws, though not psychophysical laws, may exist. Rebutting arguments from rationality, indeterminism, intensionality, etc.

McDowell, J. 1985. Functionalism and anomalous monism. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

Against Loar's functionalist reductionism: it doesn't begin to capture the normative role of rationality or the subjectivity of the mental.

McLaughlin, B.P. 1985. Anomalous monism and the irreducibility of the mental. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

A very thorough summary of Davidson's views. Highly recommended.

McLaughlin, B.P. & LePore, E. (eds) 1985. *Actions and Events*. Blackwell. 30 essays on Davidson.

McLaughlin, B.P. 1992. On Davidson's response to the charge of epiphenomenalism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Comments on Davidson 1992. Davidson can respond to critics accepting causal relevance of mental properties and still denying strict laws. Davidson misconstrues his critics' positions on supervenience.

Melchert, N. 1986. What's wrong with anomalous monism. *Journal of Philosophy* 80:265-74.

Davidson is concerned with intentional, not phenomenal states; and his characterization of these is just as physical states under a certain description. So he avoids epiphenomenalism (contra e.g. Honderich 1982).

Miller, A. 1993. Some anomalies in Kim's account of Davidson. *Southern Journal of Philosophy* 31:335-44.

Kim's version of Davidson's argument against psychophysical laws cannot work. Elucidating the notion of a constitutive principle.

Noren, S.J. 1979. Anomalous monism, events, and 'the mental'. *Philosophy and Phenomenological Research* 40:64-74.

Patterson, S.A. 1996. The anomalism of psychology. *Proceedings of the Aristotelian Society* 96:37-52.

Preyer, G. 2000. Primary reasons: From radical interpretation to a pure anomalism of the mental. *Protosociology* 14:158-179.

Rosenberg, A. 1985. Davidson's unintended attack on psychology. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

Anomalous monism implies that there aren't even heteronomic psychological generalizations, as variables can't be independently measured.

Rowlands, M. 1990. Anomalism, supervenience, and Davidson on content-individuation. *Philosophia* 295-310.

Supervenience is compatible with anomalism: biconditional laws are ruled out by the disjunctive base, and the wideness of mental states rules out one-way psychophysical laws, as there's no single property in the base.

Seager, W.E. 1981. The anomalousness of the mental. *Southern Journal of Philosophy* 19:389-401.

Elucidating Davidson's argument, focusing on the argument against strict psychophysical laws. Generalizations involve disjunctive kinds and so are heteronomic and not law-like.

Seager, W.E. 1991. Disjunctive laws and supervenience. *Analysis* 51:93-98.

Argues contra Kim that supervenience is compatible with anomalous monism: the the disjunctive generalizations aren't lawlike, as they aren't confirmed by their instances.

- Smart, J.J.C. 1985. Davidson's minimal materialism. In (B. Vermazen & M. Hintikka, eds) *_Essays on Davidson_*. Oxford University Press.
Some comments on holism, indeterminacy, anomalism, and materialism.
- Smith, P. 1982. Bad news for anomalous monism? *Analysis* 42:220-4.
Response to Honderich 1982: physical events are individuated as mental states by virtue of their causal role, so the mental is causally relevant.
- Sosa, E. 1993. Davidson's thinking causes. In (J. Heil & A. Mele, eds) *_Mental Causation_*. Oxford University Press.
- Stanton, W.L. 1983. Supervenience and psychophysical law in anomalous monism. *Pacific Philosophical Quarterly* 64:72-9.
Supervenience entails psychophysical principles, but this is compatible with anomalous monism. On what constitutes a strict psychophysical law.
- Suppes, P. 1985. Davidson's views on psychology as a science. In (B. Vermazen & M. Hintikka, eds) *_Essays on Davidson_*. Oxford University Press.
Various: physics is indeterministic and intensional, animals have beliefs, psychology has derived laws, and decision-theory doesn't need speech.
- Tiffany, E.C. 2001. The rational character of belief and the argument for mental anomalism. *Philosophical Studies* 103:258-314.
- van Gulick, R. 1980. Rationality and the anomalous nature of the mental. *Philosophy Research Archives* 7:1404.
Rationality constraints don't introduce an irreducibly normative element into intentional attributions. Rationality serves as a condition of adequacy for psychophysical theories, but it doesn't rule them out.
- Vermazen, B. & Hintikka, M. (eds) 1985. *_Essays on Davidson_*. Oxford University Press.
12 essays on Davidson, with replies.
- Walsh, D.M. 1998. Wide content individualism. *Mind* 107:625-652.
- Welshon, R. 1999. Anomalous monism and epiphenomenalism. *Pacific Philosophical Quarterly* 80:103-120.
- Yalowitz, S. 1997. Rationality and the argument for anomalous monism. *Philosophical Studies* 87:235-58.
- Yalowitz, S. 1998. Causation in the argument for anomalous monism. *Canadian Journal of Philosophy* 28:183-226.
- Zangwill, N. 1993. Supervenience and anomalous monism: Blackburn on Davidson. *Philosophical Studies* 71:59-79.

3.6 Mental Causation [see also 2.2c]

- Antony, L. 1991. The causal relevance of the mental. *Mind and Language* 6:295-327.
- Audi, R. 1993. Mental causation: Sustaining and dynamic. In (J. Heil & A. Mele, eds) *_Mental Causation_*. Oxford University Press.
- Baker, L.R. 1993. Metaphysics and mental causation. In (J. Heil & A. Mele, eds) *_Mental Causation_*. Oxford University Press.
Mental causation is incompatible with strong supervenience and causal closure of physics, as we can't distinguish high-level causes from non-causes. So reject the metaphysics and make explanation prior to causation.

Barrett, J. 1994. Rationalizing explanation and causally relevant mental properties. *Philosophical Studies* 74:77-102.

Blackburn, S. 1991. Losing your mind: Physics, identity, and folk burglar prevention. In (J. Greenwood, ed) *The Future of Folk Psychology*. Cambridge University Press.

Arguing for the causal efficacy and scientific respectability of higher-order states, such as functional-role states. To require appeal to particular physical states is to succumb to a "Tractarian" view of physical primacy.

Block, N. 1989. Can the mind change the world? In (G. Boolos, ed) *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.

Rescuing content from epiphenomenalism via functional role argument; but then functional roles aren't really causally efficacious (cf. dormitive virtue), so epi all over again? Roles vs fillers, causation vs explanation.

Block, N. 1995. Reply: Causation and two kinds of laws. In (C. Macdonald & G. Macdonald, eds) *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford University Press.

Braun, D. 1995. Causally relevant properties. *Philosophical Perspectives* 9:447-75.

Brewer, B. 1995. Compulsion by reason (Mental Causation II). *Aristotelian Society Supplement* 69:237-53.

Burge, T. 1993. Mind-body causation and explanatory practice. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Mental causation is not a real worry, but the to-do shows that materialist metaphysics has shed little light on it. It needs to be understood at the mental level. With remarks on exclusion arguments and token identity.

Crane, T. 1990. On an alleged analogy between numbers and propositions. *Analysis* 50:224-30.

How can a relation to a proposition (an abstract object) be causally efficacious? Analogy with numbers doesn't work: weight properties are only pseudo-relational, depending on units, but propositions are absolute.

Crane, T. 1992. Mental causation and mental reality. *Proceedings of the Aristotelian Society* 66:185-202.

Argues that anomalism and causal closure don't pose problems for mental causation as they are false, and that functional properties can be efficacious. States with content may be efficacious, although content itself may not be.

Crane, T. 1995. The mental causation debate (Mental causation I). *Aristotelian Society Supplement* 69:211-36.

Argues that mental causation is a deep problem for constitutive (but not identity) forms of physicalism. The only way out is to argue that it is a different variety of causation. But then what motivates physicalism?

Dretske, F. 1993. Mental events as structuring causes of behavior. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Mental events are structuring causes of behavior; biological events are triggering causes, dependent on previous mental structuring. This allows extrinsic properties to play a causal role.

Ehring, D. 1996. Mental causation, determinables, and property instances. *Nous* 30:461-80.

Hardcastle, V.G. 1998. On the matter of minds and mental causation. *Philosophy and Phenomenological Research* 58:1-25.

- Heil, J. 1992. Mentality and causality. *Topoi* 11:103-110.
On various problems with mental causation, and the relationship between psychology and philosophy.
- Honderich, T. 1993. The union theory and anti-individualism. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.
The identity theory and psychoneural correlation can't handle mental causation; only the union theory can. Anti-individualism causes problems, but should be rejected in any case.
- Horgan, T. 1989. Mental causation. *Philosophical Perspectives* 3:47-74.
How mental events are causally relevant qua mental: via an account of "qua" causation in general, using counterfactuals on "pertinently similar worlds".
- Horgan, T. 1997. Kim on mental causation and causal exclusion. *Philosophical Perspectives* 11:165-84.
- Hornsby, J. 1993. Agency and causal explanation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.
- Jackson, F. & Pettit, P. 1990. Causation and the philosophy of mind. *Philosophy and Phenomenological Research Supplement* 50:195-214.
A defense of functional role as a causally efficacious property of physical states. With application to connectionism & eliminativism.
- Jackson, F. & Pettit, P. 1990. Program explanation: A general perspective. *Analysis* 50:107-17.
- Jackson, F. 1995. Essentialism, mental properties, and causation. *Proceedings of the Aristotelian Society*.
How can content properties be causes, given that content is a matter of functional role and that functional properties are not causes? Defends a type-identity answer against various objections.
- Jackson, F. 1996. Mental causation. *Mind* 105:377-413.
A "state of the art" review paper, concentrating on problems posed by autonomy, functionalism, and externalism, and advocating a sort of identity theory. With discussion of a "map-system" view vs. a language of thought.
- Kazez, J.R. 1995. Can counterfactuals save mental causation? *Australasian Journal of Philosophy* 73:71-90.
- Kim, J. 1984. Epiphenomenal and supervenient causation. *Midwest Studies in Philosophy* 9:257-70. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).
Psychological causation, like all macrocausation, is supervenient epiphenomenal causation.
- Kim, J. 1992. The nonreductivist's trouble with mental causation. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).
Argues that nonreductive materialism implies downward causation (as the mental has more causal powers than the physical alone), and that downward causation violates the causal closure of the physical.
- Kim, J. 1992. "Downward causation" in emergentism and nonreductive physicalism. In (A. Beckermann, H. Flohr, & J. Kim, eds) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.
Argues that nonreductive materialism is just like 1930s emergentism, with the the mental contributing new causal powers, and so implies downward causation.
- Kim, J. 1993. Mental causation in a physical world. In (E. Villanueva, ed) *Science and Knowledge*. Ridgeview.

- Kim, J. 1994. 'Second-order' properties and mental causation. Manuscript.
- Kim, J. 1995. Mental causation: What? Me worry? In (E. Villanueva, ed) _Content_. Ridgeview.
- Leiter, B. & Miller, A. 1994. Mind doesn't matter yet. *Australasian Journal of Philosophy* 72:220-28.
Argues that the arguments of Fodor and LePore & Loewer don't succeed in defeating the threat of epiphenomenalism.
- Leiter, B. & Miller, A. 1998. Closet dualism and mental causation. *Canadian Journal of Philosophy* 28:161-181.
- LePore, E. & Loewer, B. 1989. More on making mind matter. *Philosophical Topics* 17:175-91.
On the problems that irreducibility -- multiple realizability, normativity, and non-supervenience -- poses for mental causation. Criticizes Kim's supervenient causation and Fodor's causal powers, and looks to "quasation".
- Macdonald, C. & Macdonald, G. 1986. Mental causes and explanation of action. *Philosophical Quarterly* 36:145-58.
- Macdonald, C. & Macdonald, G. 1991. Mental causation and nonreductive monism. *Analysis* 51:23-32.
- Macdonald, C. & Macdonald, G. 1995. How to be psychologically relevant. In (C. Macdonald & G. Macdonald, eds) _Philosophy of Psychology: Debates on Psychological Explanation_. Oxford University Press.
- Macdonald, G. 1992. The nature of naturalism. *Aristotelian Society Supplement* 66:225-44.
- Marras, A. 1994. Nonreductive materialism and mental causation. *Canadian Journal of Philosophy* 24:465-93.
- Marras, A. 1997. The causal relevance of mental properties. *Philosophia* 25:389-400.
- Marras, A. 1998. Kim's principle of explanatory exclusion. *Australasian Journal of Philosophy* 76:439-451.
- McGrath, M. 1998. Proportionality and mental causation: A fit? *Philosophical Perspectives* 12:167-176.
- McLaughlin, B.P. 1989. Type epiphenomenalism, type dualism, and the causal priority of the physical. *Philosophical Perspectives* 3:109-135.
Physical comprehensiveness and mental/physical non-reductionism don't imply mental inefficacy; nor does anomalous monism. Non-physical types can still be causal, though they must be accompanied by physical causation.
- Noordhof, P. 1997. Making the change: The functionalist's way. *British Journal for the Philosophy of Science* 48:233-??.
- Noordhof, P. 1998. Do tropes resolve the problem of mental causation? *Philosophical Quarterly* 48:221-26.
- Pettit, P. 1992. The nature of naturalism. *Aristotelian Society Supplement* 66:245-66.
On making sense of the causal efficacy of higher-level properties under naturalism. They're relevant at the program level, not quite in the way that basic properties are. With remarks on Macdonald's objections.
- Robb, D. 1997. The properties of mental causation. *Philosophical Quarterly*

187:178-94.

Robb, D. 2001. Reply to Noordhof on mental causation. *Philosophical Quarterly* 51:90-94.

Robinson, W.S. 1979. Do pains make a difference to our behavior? *American Philosophical Quarterly* 16:327-34.

On Goldman's (1969) argument that dualism and causal closure are compatible with mental causation. Goldman establishes only hypothetical necessity, not causal necessity

Searle, J.R. 1984. Intentionality and its place in nature. *Synthese* 61:3-16. Intentionality is caused by the physical, and causes. More a 1P emphasis.

Sosa, E. 1984. Mind-body interaction and supervenient causation. *Midwest Studies in Philosophy* 9:271-81.

Interactionist dualism is out, supervenient causation is in. But there are problems with mental events' causal relevance qua mental, especially for anomalous monism. Cf: a loud shot causes death, but loudness isn't relevant.

Thomasson, A. 1998. A nonreductivist solution to mental causation. *Philosophical Studies* 89:181-95.

Tuomela, R. 1998. A defense of mental causation. *Philosophical Studies* 90:1-34.

van Gulick, R. 1993. Who's in charge here? And who's doing all the work? In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

On three arguments against mental causation, from strict laws, non-local supervenience, and especially exclusion. Mental properties are stable, recurring high-level patterns with their own causal relevance.

Worley, S. 1997. Determination and mental causation. *Erkenntnis* 46:281-304.

Yablo, S. 1992. Mental causation. *Philosophical Review* 101:245-280.

Argues that mental events/properties stand to physical events/properties as determinable to determinates, solving the exclusion problem. Some mental events are *better* candidates for the cause of action than physical events.

Zangwill, N. 1996. Good old supervenience: Mental causation on the cheap. *Synthese* 106:67-101.

Argues that anomalous monism is compatible with mental causation: supervenience is necessary and sufficient for causal efficacy.

3.7 Personal Identity

3.7a Personal Identity, General

Baillie, J. 1993. Recent work on personal identity. *Philosophical Books* 34:193-206.

Baillie, J. 1997. Personal identity and mental content. *Philosophical Psychology* 10:323-33.

Brennan, A. 1982. Personal identity and personal survival. *Analysis* 42:44-50.

Brennan, A. 1984. Survival. *Synthese* 59:339-62.

Brennan, A. 1987. Discontinuity and identity. *Nous* 21:241-60.

Brennan, A. 1988. *Conditions of Identity: A Study of Identity and Survival*.

Oxford University Press.

Brooks, D.H.M. 1986. Group minds. Australasian Journal of Philosophy 64:456-70.

Carter, W. 1999. Will I be a dead person? Philosophy and Phenomenological Research 59.

Cartwright, H.M. 1987. Ruminations on an account of personal identity. In (J.J. Thomson, ed) *On Being and Saying: Essays on Honor of Richard Cartwright*. MIT Press.

Cartwright, H.M. 1993. On two arguments for the indeterminacy of personal identity. Synthese 95:241-273.

Cockburn, D. (ed) 1991. *Human Beings*. Cambridge University Press.

Coleman, S. 2000. Thought experiments and personal identity. Philosophical Studies 98:51-66.

Cowley, F. 1971. The identity of a person and his body. Journal of Philosophy 68:678-683.

Dainton, B. 1996. Survival and experience. Proceedings of the Aristotelian Society 96:17-36.

Davis, L.H. 1998. Functionalism and personal identity. Philosophy and Phenomenological Research 58:781-804.

Davis, L.H. 2001. Functionalism, the brain, and personal identity. Philosophical Studies 102:259-79.

Dennett, D.C. 1978. Where am I? In *Brainstorms*. MIT Press.

Elliot, R. 1991. Personal identity and the causal continuity requirement. Philosophical Quarterly 41:55-75.

Ganeri, J. 2000. Cross-modality and the self. Philosophy and Phenomenological Research 61:639-658.

Garrett B. 1990. Personal identity and extrinsicness. Philosophical Studies 59:177-194.

Garrett, B. 1991. Personal identity and reductionism. Philosophy and Phenomenological Research 51:361-373.

Garrett, B. 1992. Persons and values. Philosophical Quarterly 42:337-44.

Glover, J. 1988. *I: The Philosophy and Psychology of Personal Identity*. Penguin.

Hamilton, A. 1995. A new look at personal identity. Philosophical Quarterly 45:332-349.

Harris, H. (ed) 1995. *Identity*. Oxford University Press.

Harris, H. 1995. An experimentalist looks at identity. In (H. Harris, ed) *Identity*. Oxford University Press.

Hasker, W. 1999. *The Emergent Self*. Cornell University Press.

Hope, T. 1994. Personal Identity and Psychiatric Illness. Philosophy 37:131-143.

- Johnston, M. 1992. Reasons and reductionism. *Philosophical Review* 3:589-618.
- Kolak, D. & Martin, R. 1987. Personal identity and causality: Becoming unglued. *American Philosophical Quarterly*.
- Kolak, D. 1993. The metaphysics and metapsychology of personal identity: Why thought experiments matter in deciding who we are. *American Philosophical Quarterly* 30:39-50.
- Kolak, D. & Martin, R. (eds) 1991. *_Self and Identity: Contemporary Philosophical Issues_*. Macmillan.
- Madell, G. 1981. *_The Identity of the Self_*. Edinburgh University Press.
- Madell, G. 1991. Personal identity and the idea of a human being. *Philosophy* 29:127-142.
- Martin, R. 1992. Self-interest and survival. *American Philosophical Quarterly* 29:319-30.
- Matthews, S. 2000. Survival and separation. *Philosophical Studies* 98:279-303.
- McCall, C. 1990. *_Concepts of Person: An Analysis of Concepts of Person, Self, and Human Being_*. Avebury.
- Merricks, T. 2000. Perdurantism and psychological continuity. *Philosophy and Phenomenological Research* 61:195-199.
- Miri, M. 1973. Memory and personal identity. *Mind* 82:1-21.
- Nerlich, G.C. 1958. Sameness, difference, and continuity. *Analysis*.
- Noonan, H. 1989. *_Personal Identity_*. Routledge.
- Noonan, H. 1993. Chisholm, persons, and identity. *Philosophical Studies* 69:35-58.
- Nozick, R. 1981. The identity of the self. In *_Philosophical Explanations_*. Harvard University Press.
- Olson E. 1994. Is Psychology relevant to personal identity? *Australasian Journal of Philosophy* 72:173-186.
- Olson, E.T. 1997. *_The Human Animal: Personal Identity without Psychology_*. Oxford University Press.
- Olson, E. 2001. Personal identity and the radiation argument. *Analysis* 61:38-44.
- Peacocke, A. & Gillett, G. (eds) 1987. *_Persons and Personality: A Contemporary Inquiry_*. Blackwell.
- Penelhum, T. 1959. Personal identity, memory, and survival. *Journal of Philosophy*.
- Penelhum, T. 1971. The importance of self-identity. *Journal of Philosophy* 68:667-78.
- Perry, J. 1972. Can the self divide? *Journal of Philosophy* 69:463-88.
- Perry, J. (ed) 1975. *_Personal Identity_*. University of California Press.
- Perry, J. 1975. Personal identity, memory, and the problem of circularity. In (J. Perry, ed) *_Personal Identity_*. University of California Press.

- Perry, J. 1976. The importance of being identical. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Perry, J. 1978. *A Dialogue on Personal Identity and Immortality*. Hackett.
- Pogue, J.E. 1993. Identity, survival, and the reasonableness of replication. *Southern Journal of Philosophy* 31:45-70.
- Rea, M. & Silver, D. 2000. Personal identity and psychological continuity. *Philosophy and Phenomenological Research* 61:185-194.
- Rey, G. 1976. Survival. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.
- Rieber, S. 1998. The concept of personal identity. *Philosophy and Phenomenological Research* 58:581-594.
- Robert, M. 1983. Lewis's theory of personal identity. *Australasian Journal of Philosophy* 61:58-67.
- Rorty, A. (ed) 1976. *The Identities of Persons*. University of California Press.
- Shalom, A. 1985. *The Body-Mind Conceptual Framework and the Problem of Personal Identity*. Humanities Press.
- Schechtman, M. 1990. Personhood and personal identity. *Journal of Philosophy* 87:71-92.
- Shoemaker, S. 1959. Personal identity and memory. *Journal of Philosophy* 56:868-902.
- Shoemaker, S. 1970. Persons and their pasts. *American Philosophical Quarterly* 7:269-85.
- Shoemaker, S. & Swinburne, S. 1984. *Personal Identity: Great Debates in Philosophy*. Blackwell.
- Shorter, J.M. 1962. More about bodily continuity and personal identity. *Analysis* 22:79-85.
- Unger, P. 1990. *Identity, Consciousness, and Value*. Oxford University Press.
- Vesey, P. 1974. *Personal Identity: A Philosophical Analysis*. Cornell University Press.
- White, S. 1989. Metapsychological relativism and the self. *Journal of Philosophy* 86:298-323.
- Whiting, J. 1986. Friends and future selves. *Philosophical Review* 95:547-80.
- Wilkes, K.V. 1988. *Real People: Personal Identity Without Thought Experiments*. Oxford University Press.
- Williams, B. 1957. Personal identity and individuation. *Proceedings of the Aristotelian Society* 67:229-52.
- Williams, B. 1973. *Problems of the Self*. Cambridge University Press.
- Zemach, E. 1987. Looking out for number one. *Philosophy and Phenomenological Research*.

Zuboff, A. 1978. Moment universals and personal identity. *Proceedings of the Aristotelian Society* 52:141-55.

Zuboff, A. 1990. One self: The logic of experience. *Inquiry* 33:39-68.

3.7b Parfit on Personal Identity

Baillie, J. 1993. What matters in survival. *Southern Journal of Philosophy* 31:255-61.

Baillie, J. 1996. Identity, relation R, and what matters: A challenge to Derek Parfit. *Pacific Philosophical Quarterly* 77:263-267.

Beck, S. 1989. Parfit and the Russians (personal identity and moral concepts). *Analysis* 49:205-209.

Bodansky, E. 1987. Parfit on selves and their interests. *Analysis* 47:47-50.

Brennan, A.A. 1987. Survival and importance. *Analysis* 47:225-30.

Brueckner, A. 1993. Parfit on what matters in survival. *Philosophical Studies* 70:1-22.

Bushnell, D.E. 1993. Identity, psychological continuity, and rationality. *Journal of Philosophical Research* 18:15-24.

Campbell, S. 2000. Strawson, Parfit and impersonality. *Canadian Journal of Philosophy* 30:207-225.

Cassam, Q. 1993. Parfit on persons. *Proceedings of the Aristotelian Society* 93:17-37.

Chappell, T. 1995. Personal identity, R-relatedness, and the empty question argument. *Philosophical Quarterly* 45:88-92.

Chappell, T. 1998. Reductionism about persons; and what matters. *Proceedings of the Aristotelian Society* 98:41-58.

Collins, A.W. 1997. Personal identity and the coherence of q-memory. *Philosophical Quarterly* 47:73-80.

Curzer, H. 1991. An ambiguity in Parfit's theory of personal identity. *Ratio* 4:16-24.

Dancy, J. (ed). 1997. *_Reading Parfit_*. Blackwell.

Doepke, F. 1990. The practical importance of personal identity. *Logos* 83-91.

Ehring, D. 1987. Survival and trivial facts. *Analysis* 47:50-54.

Ehring, D. 1995. Personal identity and the R-relation: Reconciliation through cohabitation. *Australasian Journal of Philosophy* 73:337-346.

Fields, L. 1987. Parfit on personal identity and desert. *Philosophical Quarterly* 37:432-41.

Gillett, G. 1987. Reasoning about persons. In (A. Peacocke & G. Gillett, eds) *_Persons and Personality: A Contemporary Inquiry_*. Blackwell.

Goodenough, J.M. 1996. Parfit and the Sorites paradox. *Philosophical Studies* 2:113-20.

Haugen, D. 1995. Personal identity and concern for the future. *Philosophia*

24:481-492.

Hirsch, E. 1991. Divided minds. *Philosophical Review* 1:3-30.

Johnston, M. 1989. Fission and the facts. *Philosophical Perspectives* 3:369-97.

Korsgaard, C. 1989. Personal identity and the unity of agency: A Kantian response to Parfit. *Philosophy and Public Affairs* 18:103-31.

Lee, W. 1990. Personal identity, the temporality of agency, and moral responsibility. *Auslegung* 16:17-29.

Lewis, D. 1976. Survival and identity. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.

Madell, G. 1985. Derek Parfit and Greta Garbo. *Analysis* 45:105-9.

Maddy, P. 1979. Is the importance of identity derivative? *Philosophical Studies* 35:151-70.

Matthews, G.B. 1977. Surviving as. *Analysis* 37:53-58.

Martin, R. 1987. Memory, connecting, and what matters in survival. *Australasian Journal of Philosophy* 65:82-97.

Measor, N. 1980. On what matters in survival. *Mind* 89:406-11.

Merricks, T. 1997. Fission and personal identity over time. *Philosophical Studies* 88:163-186.

Northoff, G. 2000. Are "q-memories" empirically realistic?: A neurophilosophical approach. *Philosophical Psychology* 13:191-211.

Oaklander, L.N. 1987. Parfit, circularity, and the unity of consciousness. *Mind* 96:525-29.

Parfit, D. 1971. Personal identity. *Philosophical Review* 80:3-27.

Parfit, D. 1971. On the importance of self-identity. *Journal of Philosophy* 68:683-90.

Parfit, D. 1973. Later selves and moral principles. In (A. Montefiore, ed) *Philosophy and Personal Relations*. Routledge and Kegan Paul.

Parfit, D. 1976. Lewis, Perry, and what matters. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.

Parfit, D. 1982. Personal identity and rationality. *Synthese* 53.

Parfit, D. 1984. *Reasons and Persons*. Oxford University Press.

Parfit, D. 1995. The unimportance of identity. In (H. Harris, ed) *Identity*. Oxford University Press.

Robinson, J. 1988. Personal identity and survival. *Journal of Philosophy* 85:319-28.

Rovane, C. 1990. Branching self-consciousness. *Philosophical Review* 99:355-95.

Siderits, M. 1988. Ehring on Parfit's relation R. *Analysis* 48:29-32.

Slors, M. 2001. Personal identity, memory, and circularity: An alternative for q-memory. *Journal of Philosophy* 98:186-214.

Sprigge, T.L.S. 1988. Personal and impersonal identity. *Mind* 97:29-49.

Storl, H. 1992. The problematic nature of parfitian persons. *Personalist Forum* 8:123-31.

Stone, J. 1988. Parfit and the Buddha: Why there are no people. *Philosophy and Phenomenological Research* 48:519-32.

Wolf, S. 1986. Self-interest and interest in selves. *Ethics* 96:704-20.

3.7c Persons

Aune, B. 1994. Speaking of selves. *Philosophical Quarterly* 44:279-93.

Barresi, J. 1999. On becoming a person. *Philosophical Psychology* 12:79-98.

Bertocci, P.A. 1978. The essence of a person. *Monist* 61:28-41.

Biro, J.I. 1981. Persons as corporate entities and corporations as persons. *Nature and System* 3:173-80.

Chisholm, R.M. 1976. *Person and Object: A Metaphysical Study*. Open Court.

Dennett, D.C. 1976. Conditions of personhood. Lewis, D. 1976. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.

Dennett, D.C. 1989. The origins of selves. *Cogito* 3:163-73.

Heinimaa, M. 2000. Ambiguities in the psychiatric use of the concepts of the person: An analysis. *Philosophy, Psychiatry, and Psychology* 7:125-136.

Lowe, E.J. 1991. Real selves: Persons as a substantial kind. *Philosophy* 29:87-107.

Johnston, M. 1987. Human beings. *Journal of Philosophy* 84:59-83.

Margolis, J. 1988. Minds, selves, and persons. *Topoi* 7:31-45.

McInerney, P.K. 1998. Persons and psychological systems. *American Philosophical Quarterly* 35:179-193.

McInerney, P.K. 2000. Conceptions of persons and persons through time. *American Philosophical Quarterly* 37:121-134.

Oderberg, D. 1989. Johnston on human beings. *Journal of Philosophy* 86:137-41.

Olson, E. 1998. Human atoms. *Australasian Journal of Philosophy* 76:396-406.

Peterson, J. 1985. Persons and the problem of interaction. *Modern Schoolman* 62:131-38.

Rorty, A.O. 1976. A literary postscript: Characters, persons, selves, individuals. In (A. Rorty, ed) *The Identities of Persons*. University of California Press.

Smart, B. 1976. Synchronous and diachronous selves. *Canadian Journal of Philosophy* 6:13-33.

Strawson, P. 1958. Persons. *Minnesota Studies in the Philosophy of Science* 2:330-53.

Unger, P. 1979. I do not exist. In (G. Macdonald, ed) *Perception and Identity*. Cornell University Press.

Unger, P. 1979. Why there are no people. *Midwest Studies in Philosophy* 4:177-222.

Vincent, A. 1989. Can groups be persons? *Review of Metaphysics* 42:687-715.

Wiggins, D. 1987. The person as object of science, as subject of experience, and as locus of value. In (A. Peacocke & G. Gillett, eds) *Persons and Personality*. Blackwell.

3.7d Split Brains [see also 6.1e]

Baillie, J. 1991. Split brains and single minds. *Journal of Philosophical Research* 16:11-18.

Davis, L. 1997. Cerebral hemispheres. *Philosophical Studies* 87:207-22.

Gill, J.H. 1980. Of split brains and tacit knowing. *International Philosophical Quarterly* 20:49-58.

Gillett, G. 1986. Brain bisection and personal identity. *Mind* 95:224-9.

Greenwood, J.D. 1993. Split brains and singular personhood. *Southern Journal of Philosophy* 31:285-306.

Marks, C. 1980. *Commissurotomy, Consciousness, and Unity of Mind*. MIT Press.

Martin, R. 1995. Fission rejuvenation. *Philosophical Studies* 80:17-40.

Merricks, T. 1997. Fission and personal identity over time. *Philosophical Studies* 88:163-186.

Mills E. 1993. Dividing without reducing: Bodily fission and personal identity. *Mind* 102:37-51.

Moor, J. 1982. Split brains and atomic persons. *Philosophy of Science* 49:91-106.

Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22:396-413. Reprinted in *Mortal Questions* (Cambridge University Press, 1979).

Parfit, D. 1987. Divided minds and the nature of persons. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Puccetti, R. 1973. Brain bisection and personal identity. *British Journal for the Philosophy of Science* 24:339-55.

Puccetti, R. 1973. Multiple identity. *Personalist* 54:203-13.

Puccetti, R. 1975. The mute self: A reaction to DeWitt's alternative account of the split-brain data. *British Journal for the Philosophy of Science* 27:65-73.

Puccetti, R. 1981. The case for mental duality: Evidence from split-brain data and other considerations. *Behavioral and Brain Sciences* 4:93-123.

Puccetti, R. 1989. Two brains, two minds. *British Journal for the Philosophy of Science* 40:137-44.

Puccetti, R. 1993. Mind with a double brain. *British Journal for the Philosophy of Science* 44:675-92.

Puccetti, R. 1993. Dennett on the split-brain. *Psycology* 4(52).

Robinson, D.N. 1976. What sort of persons are hemispheres? Another look at "split-brain" man. *British Journal for the Philosophy of Science* 27:73-8.

Shaffer, J. 1977. Personal identity: The implications of brain bisection and brain transplants. *Journal of Medicine and Philosophy* 2:147-61.

Sperry, R.W. 1984. Consciousness, personal identity and the divided brain. *Neuropsychologia* 22:611-73.

3.7e Multiple Personality

Apter, A. 1991. The problem of who: Multiple personality, personal identity, and the double brain. *Philosophical Psychology* 4:219-48.

Benner, D.G., Evans, C.S. 1984. Unity and multiplicity in hypnosis, commissurotomy, and multiple personality disorder. *Journal of Mind and Behavior* 5:423-431.

Boden, M.A. 1994. Multiple personality and computational models. *Philosophy* 37:103-114.

Braude, S.E. 1991. *_First-person Plural: Multiple Personality and the Philosophy of Mind_*. Routledge.

Braude, S.E. 1996. Multiple personality disorder and moral responsibility. *Philosophy, Psychiatry, and Psychology* 3:37-54.

Clark, S.R.L. 1991. How many selves make me? *Philosophy* 29:213-33.

Clark, S.R.L. 1996. Minds, memes, and multiples. *Philosophy, Psychiatry, and Psychology* 3:21-28.

Flanagan, O. 1994. Multiple identity, character transformation, and self-reclamation. In (G. Graham & G. Stephens, eds) *_Philosophical Psychopathology_*. MIT Press.

Gillett, G. 1986. Multiple personality and the concept of a person. *New Ideas in Psychology* 4:173-84.

Gillett, G. 1997. A discursive account of multiple personality disorder. *Philosophy, Psychiatry, and Psychology* 4:213-22.

Hacking, I. 1991. Two souls in one body. *Critical Inquiry* 17:838-67.

Hacking, I. 1995. *_Rewriting the Soul: Multiple Personality and the Sciences of Memory_*. Princeton University Press.

Humphrey, N. & Dennett, D.C. 1989. Speaking for ourselves. *Raritan* 9:68-98.

Kolak, D. 1993. Finding our selves: Identification, identity, and multiple personality. *Philosophical Psychology* 6:363-86.

Lizza, J.P. 1993. Multiple personality and personal identity revisited. *British Journal for the Philosophy of Science* 44:263-274.

Matthews, S. 1998. Personal identity, multiple personality disorder, and moral personhood. *Philosophical Psychology* 11:67-88.

Radden, J. 1996. *_Divided Minds and Successive Selves: Ethical Issues in Disorders of Identity and Personality_*. MIT Press.

Wilkes, K.V. 1981. Multiple personalty and personal identity. *British Journal for the Philosophy of Science* 32:331-48.

Wilkes, K.V. 1991. How many selves make me? *Philosophy* 66:235-43.

3.8 Free Will -----

Albritton, R. 1985. Freedom of the will and freedom of action. *Proceedings and Addresses of the American Philosophical Association* 59:239-51.

Anscombe, G.E.M. 1976. 'Soft' determinism. In (G. Ryle, ed) *Contemporary Aspects of Philosophy*. Oriel Press.

Audi, R. 1974. Moral responsibility, freedom, and compulsion. *American Philosophical Quarterly* 11:1-14.

Ayer, A.J. 1980. Free will and rationality. In (Z. van Straaten, ed) *Philosophical Subjects*. Oxford University Press.

Ayers, M. 1968. *The Refutation of Determinism*. Methuen.

Berofsky, B. (ed) 1966. *Free Will and Determinism*. Harper and Row.

Berofsky, B. 1971. *Determinism*. Princeton University Press.

Berofsky, B. 1987. *Freedom from Necessity: The Metaphysical Basis of Responsibility*. Routledge.

Bishop, J. 1993. Compatibilism and the free will defense. *Australasian Journal of Philosophy* 71:104-20.

Blumenfeld, D. 1971. The principle of alternate possibilities. *Journal of Philosophy* 67:339-44.

Blumenfeld, D. 1988. Freedom and mind control. *American Philosophical Quarterly* 25:215-27.

Campbell, C.A. 1951. Is "free will" a pseudoproblem? *Mind* 60:441-65.

Churchland, P.S. 1981. Is determinism self-refuting? *Mind* 90:99-101.

Clarke, R. 1992. Free will and the conditions of moral responsibility. *Philosophical Studies* 66:53-72.

Clarke, R. 1993. Toward a credible agent-causal account of free will. *Nous* 27:191-203.

Clarke, R. 2000. Modest libertarianism. *Philosophical Perspectives* 14:21-46.

Crisp, T. & Warfield, T. 2000. The irrelevance of indeterministic counterexamples to principle beta. *Philosophy & Phenomenological Research* 61:173-185.

Dennett, D.C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.

Double, R. 1989. Puppeteers, hypnotists, and neurosurgeons. *Philosophical Studies* 56:163-73.

Double, R. 1991. *The Non-Reality of Free Will*. Oxford University Press.

Double, R. 1991. Determinism and the experience of freedom. *Pacific Philosophical Quarterly* 72:1-8.

- Double, R. 1992. How rational must free will be? *Metaphilosophy* 23:268-78.
- Double, R. 1994. How to frame the free will problem. *Philosophical Studies* 75:149-72.
- Double, R. 1996. *Metaphilosophy and Free Will*. Oxford University Press.
- Duggan, T. & Gert, B. 1979. Free will as the ability to will. *Nous* 13:197-217.
- Dworkin, G. (ed) 1970. *Determinism, Free Will, and Moral Responsibility*. Prentice-Hall.
- Eccles, J. 1976. Brain and free will. In (G. Globus, ed) *Consciousness and the Brain*. Plenum Press.
- Fischer, J.M. 1982. Responsibility and control. *Journal of Philosophy* 79:24-40.
- Fischer, J.M. & Ravizza, M. 1992. When the will is free. *Philosophical Perspectives* 6:423-51.
- Fischer, J.M. 1994. *The Metaphysics of Free Will*. Blackwell.
- Fischer, J.M. & Ravizza, M. 1996. Free will and the modal principle. *Philosophical Studies* 3:213-30.
- Fowler, C. 1996. A pragmatic defense of free will. *Journal of Value Inquiry* 30:247-60.
- Frankfurt, H. 1969. Alternate possibilities and moral responsibility. *Journal of Philosophy* 65:829-39.
- Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68:5-20.
- Furlong, F.W. 1981. Determinism and free will: Review of the literature. *American Journal of Psychiatry* 138:435-39.
- Garson, J.W. 1995. Chaos and free will. *Philosophical Psychology* 8:365-74.
- Ginet, C. 1980. The conditional analysis of freedom. In (P. van Inwagen, ed) *Time and Cause: Essays Presented to Richard Taylor*. Reidel.
- Ginet, C. 1983. In defense of incompatibilism. *Philosophical Studies* 44:391-400.
- Ginet, C. 1989. Reasons explanation of action: An incompatibilist account. *Philosophical Perspectives* 3.
- Ginet, C. 1990. *On Action*. Cambridge University Press.
- Goldman, A. 1968. Actions, predictions, and books of life. *American Philosophical Quarterly*.
- Goldman, A. 1969. The compatibility of mechanism and purpose. *Philosophical Review* 78:468-82.
- Griffiths, A.P. 1989. Is free will incompatible with something or other? *Philosophy* 24:101-19.
- Greenspan, P.S. 1978. Behavior control and freedom of action. *Philosophical Review* 87:225-40.

Greenspan, P.S. 1993. Free will and the genome project. *Philosophy and Public Affairs* 22:31-43.

Heller, M. 1996. The mad scientist meets the robot cats: Compatibilism, kinds, and counterexamples. *Philosophy and Phenomenological Research* 56:333-37.

Honderich, T. (ed) 1973. *Essays on Freedom of Action*. Routledge and Kegan Paul.

Honderich, T. 1988. *A Theory of Determinism*. Oxford University Press.

Honderich, T. 1993. *How Free Are You?* Oxford University Press.

Horgan, T. 1985. Compatibilism and the consequence argument. *Philosophical Studies* 47:339-56.

Hospers, J. 1950. Meaning and free will. *Philosophy and Phenomenological Research* 10:307-30.

Howard, G.S. 1993. Steps toward a science of free will. *Counseling and Values* 37:116-28.

Kane, R. 1985. *Free Will and Values*. SUNY Press.

Kane, R. 1989. Two kinds of incompatibilism. *Philosophy and Phenomenological Research* 69:219-54.

Kane, R. 1994. Free will: The illusive ideal. *Philosophical Studies* 75:25-60.

Kane, R. 1996. *The Significance of Free Will*. Oxford University Press.

Kane, R. 2000. The dual regress of free will and the role of alternative possibilities. *Philosophical Perspectives* 14:57-80.

Kapitan, T. 1986. Deliberation and the presumption of open alternatives. *Southern Journal of Philosophy* 40:230-51.

Kapitan, T. 1991. Ability and cognition: A defense of compatibilism. *Philosophical Studies* 63:231-43.

Kapitan, T. 1996. Modal principles in the metaphysics of free will. *Philosophical Perspectives* 10:419-45.

Kapitan, T. 2000. Autonomy and manipulated freedom. *Philosophical Perspectives* 14:81-104.

Kenny, A. 1976. *Will, Freedom, and Power*. Blackwell.

Kenny, A. 1978. *Free Will and Responsibility*. Routledge.

Klein, M. 1990. *Determinism, blameworthiness, and deprivation*. Oxford University Press.

Ladd, J. 1952. Free will and voluntary action. *Philosophy and Phenomenological Research* 12:392-405.

Lahav, R. 1991. Between pre-determinism and arbitrariness: A Bergsonian approach to free will. *Southern Journal of Philosophy* 29:487-99.

Lamb, J.W. 1977. On a proof of incompatibilism. *Philosophical Review* 86:20-35.

Lamb, J.W. 1993. Evaluative compatibilism and the principle of alternate possibilities. *Journal of Philosophy* 90:517-27.

- Lehrer, K. (ed) 1966. *Freedom and Determinism*. Random House.
- Lehrer, K. 1966. An empirical disproof of determinism. In (K. Lehrer, ed) *Freedom and Determinism*. Random House.
- Lehrer, K. 1976. 'Can' in theory and practice: A possible worlds analysis. In (M. Brand & D. Walton, eds) *Action Theory*. Reidel.
- Locke, D. 1975. Three concepts of free action. *Aristotelian Society Supplement* 75:95-112.
- Lucas, J.R. 1970. *The Freedom of the Will*. Oxford University Press.
- Machina, K. 1994. Challenges for compatibilism. *American Philosophical Quarterly* 31:213-22.
- Margenau, H. 1931. The uncertainty principle and free will. *Science*.
- McCall, S. 1984. Freedom defined as the power to decide. *American Philosophical Quarterly* 21:329-38.
- Morden, M. 1990. Free will, self-causation, and strange loops. *Australasian Journal of Philosophy* 68:59-73.
- Morgenbesser, S. & Walsh, J.J. (eds) 1962. *Freedom and Responsibility*. Prentice-Hall.
- Narveson, J. 1977. Compatibilism defended. *Philosophical Studies* 32:83-7.
- Neely, W. 1974. Freedom and desire. *Philosophical Review* 83:32-54.
- O'Connor, D.J. 1971. *Free Will*. Anchor Books.
- O'Connor, T. 1993. Indeterminism and free agency: Three recent views. *Philosophy and Phenomenological Research* 53:499-26.
- O'Connor, T. (ed) 1995. *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. Oxford University Press.
- O'Leary-Hawthorne, J. & Pettit, P. 1996. Strategies for free will compatibilists. *Analysis* 56:191-201.
- O'Shaughnessy, B. 1980. *The Will: A Dual Aspect Theory*. Cambridge University Press.
- Pereboom, D. 2000. Alternative possibilities and causal histories. *Philosophical Perspectives* 14:119-138.
- Perszyk, K.J. 1999. Compatibilism and the free will defence: A reply to Bishop. *Australasian Journal of Philosophy* 77:92-105.
- Popper, K. 1983. Is determinism self-refuting? *Mind* 92:103-4.
- Ravizza, M. 1994. Semi-compatibilism and the transfer of non-responsibility. *Philosophical Studies* 75:61-93.
- Rowe, W. 1987. Two concepts of freedom. *Proceedings and Addresses of the American Philosophical Association* 61:43-64.
- Rychlak, J.F. 1976. Can psychology be objective about free will? *Philosophical Psychologist* 10:2-9. Revised version in *New Ideas in Psychology* 1:213-29, 1983.

- Rychlak, J.F. 1994. Four kinds of determinism and "free will": A response to Viney and Crosby. *New Ideas in Psychology* 12:143-46.
- Rychlak, J.F. 1994. Is free will a process or a content: Both? neither? Are we free to take a position on this question? *Journal of Theoretical and Philosophical Psychology* 14:62-72.
- Sappington, A.A. 1990. Recent psychological approaches to the free will versus determinism controversy. *Psychological Bulletin* 108:19-29.
- Searle, J. 2000. Consciousness, free action and the brain. *Journal of Consciousness Studies* 7:3-22.
- Settle, T. 1993. How determinism refutes compatibilism. *Religious Studies* 29:353-62.
- Slife, B.D. 1994. Free will and time: That "stuck" feeling. *Journal of Theoretical and Philosophical Psychology* 14:1-12.
- Slote, M.A. 1969. Free will, determinism, and the theory of important criteria. *Inquiry* 12:317-38.
- Slote, M. 1980. Understanding free will. *Journal of Philosophy* 77:136-51.
- Slote, M. 1982. Selective necessity and the free will problem. *Journal of Philosophy* 74:5-24.
- Spence, S.A. 1996. Free will in the light of neuropsychiatry. *Philosophy, Psychiatry, and Psychology* 3:75-90.
- Stampe, D.W. & Gibson, M.I. 1992. Of one's own free will. *Philosophy and Phenomenological Research* 52:529-56.
- Strawson, G. 1986. *Freedom and Belief*. Oxford University Press.
- Stump, E. & Fischer, J. 2000. Transfer principles and moral responsibility. *Philosophical Perspectives* 14:47-56.
- Thorp, J. 1980. *Free Will: A Defense Against Neurophysiological Determinism*. Routledge.
- van Inwagen, P. 1975. The incompatibility of free will and determinism. *Philosophical Studies* 27:185-99.
- van Inwagen, P. 1978. Ability and responsibility. *Philosophical Review* 87:201-24.
- van Inwagen, P. 1983. *An Essay on Free Will*. Oxford University Press.
- van Inwagen, P. 1989. When is the will free? *Philosophical Perspectives* 3.
- van Inwagen, P. 1994. When the will is not free. *Philosophical Studies* 75:95-113.
- Van Inwagen, P. 2000. Free will remains a mystery. *Philosophical Perspectives* 14:1-20.
- Vesey, G. 1989. Responsibility and free will. *Philosophy* 24:85-100.
- Vihvelin, K. 2000. Freedom, foreknowledge, and the principle of alternate possibilities. *Canadian Journal of Philosophy* 30:1-23.
- Vihvelin, K. 2000. Libertarian compatibilism. *Philosophical Perspectives*

14:139-166.

Viney, D.W. & Crosby, D.A. 1994. Free will in process perspective. *New Ideas in Psychology* 12:129-41.

Waller, B.N. 1989. Uneven starts and just deserts (fatalism and free will). *Analysis* 49:209-13.

Waller, B. 1990. *Freedom without Responsibility*. Temple University Press.

Warfield, T. 2000. Causal determinism and human freedom are incompatible: A new argument for incompatibilism. *Philosophical Perspectives* 14:167-180.

Watson, G. (ed) 1982. *Free Will*. Oxford University Press.

Westcott, M.R. 1977. Free will: An exercise in metaphysical truth or psychological consequences. *Canadian Psychological Review* 18:249-63.

Widerker, D. 2000. Frankfurt's attack on the principle of alternative possibilities: A further look. *Philosophical Perspectives* 14:181-202.

Williams, C. 1980. *Free Will and Determinism: A Dialogue*. Hackett.

Wilton, R. 2000. *Consciousness, Free Will, and the Explanation of Human Behavior*. E. Mellen Press.

Wolf, S. 1980. Asymmetrical freedom. *Journal of Philosophy* 77:151-66.

Wolf, S. 1981. The importance of free will. *Mind* 90:366-78.

Wolf, S. 1990. *Freedom within Reason*. Oxford University Press.

Yaffe, G. 2000. Free will and agency at its best. *Philosophical Perspectives* 14:203-230.

Zagzebski, L. 2000. Does libertarian freedom require alternate possibilities? *Philosophical Perspectives* 14:231-248.

Zimmerman, D. 1994. Acts, omissions, and semi-compatibilism. *Philosophical Studies* 73:209-23.

3.9 The Problem of Other Minds

Alexander, P. 1959. Other people's experiences. *Proceedings of the Aristotelian Society*.

Aune, B. 1961. The problem of other minds. *Philosophical Review*.

Austin, J. 1946. Other minds. *Aristotelian Society Supplement* 20:148-87.

Ayer, A.J. 1953. One's knowledge of other minds. *Theoria*.

Ayer, A.J. 1956. *The Problem of Knowledge*. Harmondsworth.

Baron-Cohen, S., Tager-Flusberg, H., Cohen, D.J. 1994. *Understanding Other Minds: Perspectives from Autism*. Oxford University Press.

Buck, R. 1962. Non-other minds. In (R. Butler, ed) *Analytic Philosophy*. Barnes and Noble.

Buford, T.O. 1970. *Essays on Other Minds*. University of Illinois Press.

Castaneda, H. 1962. Criteria, analogy, and knowledge of other minds. *Journal of Philosophy*.

- Duhrssen, A. 1963. Philosophic alienation and the problem of other minds. *Philosophical Review*.
- Everett, T. 2000. Other voices, other minds. *Australasian Journal of Philosophy* 78:213-222.
- Feigl, H. 1959. Other minds and the egocentric predicament. *Journal of Philosophy* 56:980-87.
- Gallagher, K. 1964. Intersubjective knowledge. In (Sheed & Ward, eds) *The Philosophy of Knowledge*.
- Glennan, S.S. 1995. Computationalism and the problem of other minds. *Philosophical Psychology* 8:375-88.
- Hampshire, S. 1952. The analogy of feeling. *Mind* 61:1-12.
- Harnad, S. 1991. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* 1:43-54.
On the Total Turing Test (full behavioral equivalence) as a test for mind.
- Hyslop, A. 1976. Other minds as theoretical entities. *Australasian Journal of Philosophy* 54:158-61.
- Hyslop, A. 1995. *Other Minds*. Kluwer.
- Jones, J.R. 1950. Our knowledge of other persons. *Philosophy* 25.
- Jorgensen, J. 1949. Remarks concerning the concept of mind and the problem of other people's minds. *Theoria*.
- Kurthen, M. Moskopp, D., Linke, D.B. & Reuter, B.M. 1991. The locked-in syndrome and the behaviorist epistemology of other minds. *Theoretical Medicine* 12:69-79.
- Lenman, J. 1994. Beliefs about other minds: A pragmatic justification. *American Philosophical Quarterly* 31:223-34.
- Locke, D. 1968. *Myself and Others: A Study in our Knowledge of Minds*. Oxford University Press.
- Malcolm, N. 1958. Knowledge of other minds. *Journal of Philosophy*.
- Melnyk, A. 1994. Inference to the best explanation and other minds. *Australasian Journal of Philosophy* 4:482-91.
- Mellor, W.W. 1956. Three problems about other minds. *Mind* 65:200-217.
- Morick, H. (ed) 1967. *Wittgenstein and the Problem of Other Minds*. Humanities Press.
- Narveson, A.H. 1966. Evidential necessity and other minds. *Mind* 75.
- Pap, A. 1951. Other minds and the principle of verifiability. *Revue Internationale de Philosophie* 5:280-306.
- Peacocke, C. 1984. Consciousness and other minds. *Aristotelian Society Supplement* 58:97-117.
- Plantinga, A. 1966. Induction and other minds. *Review of Metaphysics* 19:441-61.
- Plantinga, A. 1967. *God and Other Minds*. Cornell University Press.

Plantinga, A. 1968. Induction and other minds II. Review of Metaphysics 12:524-33.

Price, H.H. 1938. Our evidence for the existence of other minds. Philosophy 13:425-56.

Sagal, P. & Borg, G. 1993. The range principle and the problem of other minds. British Journal for the Philosophy of Science 44:477-91.

Slote, M. 1966. Induction and other minds. Review of Metaphysics 20:341-60.

Sober, E. 2000. Evolution and the problem of other minds. Journal of Philosophy 97:365-387.

Spencer, W. 1930. _Our Knowledge of Other Minds_. Yale University Press.

Sprigge, T.L.S. 1992. Ayer on other minds. In (L. Hahn, ed) _The Philosophy of A.J. Ayer_. Open Court.

Thalberg, I. 1969. Other times, other places, other minds. Philosophical Studies 20.

Thomson, J.F. 1951. The argument from analogy and the problem of other minds. Mind 60:336-50.

Weinberg, J. 1946. Our knowledge of other minds. Philosophical Review 60.

Wisdom, J. 1946. Other minds. Aristotelian Society Supplement 20:122-47.

Wisdom, J. 1968. _Other Minds_. University of California Press.

Zemach, E. 1966. Sensations, raw feels, and other minds. Review of Metaphysics 20:317-40.

Ziff, P. 1965. The simplicity of other minds. Journal of Philosophy 42:575-84.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 4: Philosophy of Artificial Intelligence [565] =====

4.1 Can Machines Think? [224]

4.1a The Turing Test [51]

4.1b Godelian Arguments (Lucas, Penrose) [59]

4.1c The Chinese Room (Searle) [58]

4.1d Machine Consciousness [33]

4.1e Machine Thought, Misc [23]

4.2 Computation and Representation [44]

4.2a Symbols and Symbol Systems [9]

4.2b Computational Semantics [16]

4.2c Implicit/Explicit Representation [8]

4.2d AI without Representation? [6]

4.2e Miscellaneous [6]

4.3 Philosophy of Connectionism [150]

4.3a Connectionism and Compositionality (Fodor/Pylyshyn) [43]

4.3b Representation in Connectionism [20]

4.3c Connectionism and Eliminativism [16]

4.3d The Connectionist/Classical Debate [20]

4.3e Subsymbolic Computation (Smolensky) [8]

- 4.3f Philosophy of Connectionism, Misc [36]
- 4.3g Foundational Empirical Issues [12]
- 4.4 Dynamical Systems [22]
- 4.5 Foundational Questions in AI [51]
- 4.5a The Nature of AI [7]
- 4.5b Levels of Analysis (Marr, etc) [12]
- 4.5c The Frame Problem [14]
- 4.5d Methodology [19]
- 4.6 Computationalism in Cognitive Science [33]
- 4.7 Computation and Physical Systems [18]
- 4.8 Philosophy of AI, Misc [17]

4.1 Can Machines Think?

4.1a The Turing Test

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59:433-60.

Proposes the Imitation game (Turing test) as a test for intelligence: If a machine can't be told apart from a human in a conversation over a teletype, then that's good enough. With responses to various objections.

Alper, G. 1990. A psychoanalyst takes the Turing test. *Psychoanalytic Review* 77:59-68.

Barresi, J. 1987. Prospects for the Cyberiad: Certain limits on human self-knowledge in the cybernetic age. *Journal for the Theory of Social Behavior* 17:19-46.

Block, N. 1981. Psychologism and behaviorism. *Philosophical Review* 90:5-43.

A look-up table could pass the Turing test, and surely isn't intelligent.

The TT errs in testing behavior and not mechanisms. A nice, thorough paper.

Bringsjord, S. 2001. Creativity, the Turing test, and the (better) Lovelace test. *Minds & Machines* 11:3-27.

Clark, T. 1992. The Turing test as a novel form of hermeneutics.

International Studies in Philosophy 24:17-31.

Copeland, B.J. 2000. The Turing test. *Minds and Machines* 10:519-539.

Crawford, C. 1994. Notes on the Turing test. *Communications of the Association for Computing Machinery* 37:13-15.

Crockett, L. 1994. The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence. Ablex.

Davidson, D. 1990. Turing's test. In (K. Said, ed) Modelling the Mind. Oxford University Press.

Dennett, D.C. 1984. Can machines think? In (M. Shafto, ed) How We Know. Harper & Row.

Defending the Turing test as a good test for intelligence.

Erion, G.J. 2001. The Cartesian test for automatism. *Minds and Machines* 1:29-39.

French, R.M. 1990. Subcognition and the limits of the Turing test. *Mind* 99:53-66.

The Turing Test is too hard, as it requires not intelligence but human intelligence. Any machine could be unmasked through careful questioning, but this wouldn't mean that the machine was unintelligent.

French, R.M. 1995. Refocusing the debate on the Turing Test: A response. *Behavior and Philosophy* 23:59-60.
Response to Jacquette 1993.

Gunderson, K. 1964. The imitation game. *Mind* 73:234-45.
The Turing test is not broad enough: there's much more to thought than the ability to play the imitation game.

Harnad, S. 1991. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* 1:43-54.
On the Total Turing Test (full behavioral equivalence) as a test for mind.

Harnad, S. 1994. Levels of functional equivalence in reverse bioengineering: The Darwinian Turing test for artificial life. *Artificial Life* 1(3).

Harnad, S. 1999. Turing on reverse-engineering the mind. *Journal of Logic, Language, and Information*.

Hauser, L. 1993. Reaping the whirlwind: Reply to Harnad's "Other bodies, other minds". *Minds and Machines* 3:219-37.

Hauser, L. 2001. Look who's moving the goal posts now. *Minds and Machines* 11:41-51.

Hayes, P. & Ford, K. 1995. Turing test considered harmful. *_Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence_* 1:972-77.

Hofstadter, D.R. 1981. A coffee-house conversation on the Turing test. *Scientific American*.
A dialogue on the Turing test.

Jacquette, D. 1993. Who's afraid of the Turing test? *Behavior and Philosophy* 20:63-74.
Defending the Turing test against French 1990. Turing did not intend the test to provide a *necessary* condition for intelligence.

Jacquette, D. 1993. A Turing test conversation. *Philosophy* 68:231-33.

Karelis, C. 1986. Reflections on the Turing test. *Journal for the Theory of Social Behavior* 16:161-72.

Lee, E.T. 1996. On the Turing test for artificial intelligence. *Kybernetes* 25:61.

Leiber, J. 1989. Shanon on the Turing test. *Journal of Social Behavior*.

Leiber, J. 1995. On Turing's Turing Test and why the matter matters. *Synthese* 104:59-69.
Turing's test is neutral about the structure of the machine that passes it, but it must be practical and reliable (thus excluding Searle's and Block's counterexamples).

Mays, W. 1952. Can machines think? *Philosophy* 27:148-62.

Michie, D. 1993. Turing's test and conscious thought. *Artificial Intelligence* 60:1-22. Reprinted in (P. Millican & A. Clark, eds) *_Machines and Thought_*. Oxford University Press.

Millar, P. 1973. On the point of the Imitation Game. *Mind* 82:595-97.

Moor, J.H. 1976. An analysis of Turing's test. *Philosophical Studies* 30:249-257.

The basis of the Turing test is not an operational definition of thinking, but rather an inference to the best explanation.

Moor, J.H. 1978. Explaining computer behavior. *Philosophical Studies* 34:325-7.

Reply to Stalker 1978: Mechanistic and mentalistic explanations are no more incompatible than program-based and physical explanations.

Moor, J.H. 2001. The status and future of the Turing test. *Minds and Machines* 11:77-93.

Piccinini, G. 2000. Turing's rules for the imitation game. *Minds and Machines* 10:573-582.

Purthill, R. 1971. Beating the imitation game. *Mind* 80:290-94.

Rankin, T.L. 1987. The Turing paradigm: A critical assessment. *Dialogue* 29:50-55.

Some obscure remarks on lying, imitation, and the Turing test.

Richardson, R.C. 1982. Turing tests for intelligence: Ned Block's defense of psychologism. *Philosophical Studies* 41:421-6.

A weak argument against Block: input/output function doesn't guarantee a capacity to respond sensibly.

Rosenberg, J. 1982. Conversation and intelligence. In (B. de Gelder, ed) *_Knowledge and Representation_*. Routledge & Kegan Paul.

Sampson, G. 1973. In defence of Turing. *Mind* 82:592-94.

Saygin, A.P., Cicekli, I. & Akman V. 2000. Turing test: 50 years later. *Minds and Machines* 10:463-518.

Schweizer, P. 1998. The truly total Turing Test. *Minds and Machines* 8:263-272.

Shanon, B. 1989. A simple comment regarding the Turing test. *Journal for the Theory of Social Behavior* 19:249-56.

The Turing test presupposes a representational/computational framework for cognition. Not all phenomena can be captured in teletype communication.

Shieber, S.M. 1994. Lessons from a restricted Turing test. *Communications of the Association for Computing Machinery* 37:70-82.

Stalker, D.F. 1978. Why machines can't think: A reply to James Moor. *Philosophical Studies* 34:317-20.

Contra Moor 1976: The best explanation of computer behavior is mechanistic, not mentalistic.

Sterrett, S.G. 2000. Turing's two tests for intelligence. *Minds and Machines* 10:541-559.

Stevenson, J.G. 1976. On the imitation game. *Philosophia* 6:131-33.

Traiger, S. 2000. Making the right identification in the Turing test. *Minds and Machines* 10:561-572.

Watt, S. 1996. Naive psychology and the inverted Turing test. *Psycoloquy* 7(14).

Whitby, B. 1996. The Turing test: AI's biggest blind alley? In (P. Millican & A. Clark, eds) *_Machines and Thought_*. Oxford University Press.

Zdenek, S. 2001. Passing Loebner's Turing test: A case of conflicting discourse

functions. Minds & Machines 11:53-76.

4.1b Godelian arguments (Lucas, Penrose)

Benacerraf, P. 1967. God, the Devil, and Godel. Monist 51:9-32.

Discusses and sharpens Lucas's arguments. Argues that the real consequence is that if we are Turing machines, we can't know which.

Bowie, G. 1982. Lucas' number is finally up. Journal of Philosophy Logic, 11:279-85.

Lucas's very Godelization procedure makes him inconsistent, unless he has an independent way to see if any TM is consistent, which he doesn't.

Boyer, D. 1983. J.R. Lucas, Kurt Godel, and Fred Astaire. Philosophical Quarterly 33:147-59.

Remarks on the various ways in which Lucas and a machine might be said to "prove" anything, and the ways in which a machine might simulate Lucas. The argument has all sorts of level confusions, and a bit of circularity.

Chari, C. 1963. Further comments on minds, machines and Godel. Philosophy 38:175-8.

Can't reduce the lawless creative process to computation.

Chalmers, D.J. 1996. Minds, machines, and mathematics. Psyche 2:11-20.

Chihara, C. 1972. On alleged refutations of mechanism using Godel's incompleteness results. Journal of Philosophy 64:507-26.

An analysis of the Lucas/Benacerraf argument. On various senses in which a machine might come to know its own program.

Coder, D. 1969. Godel's theorem and mechanism. Philosophy 44:234-7.

Only mathematicians understand Godel, so Lucas's argument isn't general; and Turing machines can go wrong. Weak.

Dennett, D.C. 1978. The abilities of men and machines. In *_Brainstorms_*. MIT Press.

There is no unique TM which we are -- there could be many.

Edis, T. 1998. How Godel's theorem supports the possibility of machine intelligence. Minds and Machines 8:251-262.

Feferman, S. 1996. Penrose's Godelian argument. Psyche 2:21-32.

Gaifman, H. 2000. What Godel's incompleteness result does and does not show. Journal of Philosophy 97:462-471.

George, F. 1962. Minds, machines and Godel: Another reply to Mr. Lucas. Philosophy 37:62-63.

Lucas's argument applies only to deductive machines, not inductive ones.

George, A. & Velleman, D.J. 2000. Leveling the playing field between mind and machine: A reply to McCall. Journal of Philosophy 97:456-452.

Good, I.J. 1967. Human and machine logic. British Journal for the Philosophy of Science 18:145-6.

Even humans can't Godelize forever. On ordinals and transfinite counting.

Good, I.J. 1969. Godel's theorem is a red herring. British Journal for the Philosophy of Science 19:357-8.

Rejoinder to Lucas 1967: the role of consistency; non-constructible ordinals.

Grush, R. & Churchland, P. 1995. Gaps in Penrose's toiling. In (T. Metzinger,

ed) Conscious Experience. Ferdinand Schoningh.

Hanson, W. 1971. Mechanism and Godel's theorem. *British Journal for the Philosophy of Science* 22:9-16.

An analysis of Benacerraf 1967. Benacerraf's "paradox" is illusory; there are no strong consequences of Godel's theorem for mechanism.

Hofstadter, D.R. 1979. Godel, Escher, Bach: An Eternal Golden Braid. Basic Books.

Contra Lucas: we can't Godelize forever; and we're not formal on top level.

Hutton, A. 1976. This Godel is killing me. *Philosophia* 3:135-44.

Gives a statistical argument to the effect that we cannot know that we are consistent; so the Lucas argument cannot go through.

Irvine, A.D. 1983. Lucas, Lewis, and mechanism -- one more time. *Analysis* 43:94-98.

Contra Lewis 1979, Lucas can derive the consistency of M even without the premise that he is M. Hmm.

Hadley, R.F. 1987. Godel, Lucas, and mechanical models of mind. *Computational Intelligence* 3:57-63.

A nice analysis of Lucas's argument and the circumstances under which a machine might prove another's Godel sentences. There's no reason to believe that machines and humans are different here.

Jacquette, D. 1987. Metamathematical criteria for minds and machines. *Erkenntnis* 27:1-16.

A machine will fail a Turing test if it's asked about Godel sentences.

King, D. 1996. Is the human mind a Turing machine? *Synthese* 108:379-89.

Kirk, R. 1986. Mental machinery and Godel. *Synthese*.

Lucas's argument fails, as theorems by humans don't correspond to outputs of their formal systems.

Lewis, D. 1969. Lucas against mechanism. *Philosophy* 44:231-3.

Lucas needs a rule of inference from sentences to their consistency, yielding Lucas arithmetic. No machine can prove all of Lucas arithmetic, but there's no reason to suppose humans can either, as the rule is infinitary.

Lewis, D. 1979. Lucas against mechanism II. *Canadian Journal of Philosophy* 9:373-6.

Reply to Lucas 1970: the dialectical argument fails, as the human's output depends on the premise that it is the machine (to derive M's consistency). With a similar premise, the machine itself can do equally well.

Lucas, J.R. 1961. Minds, machines and Godel. *Philosophy* 36:112-127.

Humans can Godelize any given machine, so we're not a machine.

Lucas, J.R. 1967. Human and machine logic: a rejoinder. *British Journal for the Philosophy of Science* 19:155-6.

Reply to Good 1967: a human can trump any given machine, so the human is not the machine, whether or not the human is superior across the board.

Lucas, J.R. 1968. Satan stultified: A rejoinder to Paul Benacerraf. *Monist* 52:145-58.

Benacerraf 1967 is empty and omega-inconsistent. Reply to arguments based on difficulty of seeing consistency (e.g. Putnam). Fallacious but engaging.

Lucas, J.R. 1971. Metamathematics and the philosophy of mind: A rejoinder. *Philosophy of Science* 38:310-13.

- Lucas, J.R. 1970. Mechanism: A rejoinder. *Philosophy* 45:149-51.
Response to Lewis 1969 and Coder 1969. Lewis misses the dialectical nature of the argument.
- Lucas, J.R. 1970. *The Freedom of the Will*. Oxford University Press.
- Lucas, J.R. 1976. This Godel is killing me: A rejoinder. *Philosophia* 6:145-8.
Contra Hutton, we know -- even if fallibly -- that we are consistent.
- Lucas, J.R. 1984. Lucas against mechanism II: A rejoinder. *Canadian Journal of Philosophy* 14:189-91.
Reply to Lewis 1979.
- Lucas, J.R. 1996. Mind, machines and Godel: A retrospect. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.
Addresses all the counterarguments. Fun.
- Lyngzeidetsen, A.E. & Solomon, M.K. 1994. Abstract complexity theory and the mind-machine problem. *British Journal for the Philosophy of Science* 45:549-54.
- Lyngzeidetsen, A. 1990. Massively parallel distributed processing and a computationalist foundation for cognitive science. *British Journal for the Philosophy of Science* 41.
A Connection Machine might escape the Lucas argument. Bizarre.
- Martin, J. & Engleman, K. 1990. The mind's I has two eyes. *Philosophy* 510-16.
Contra Hofstadter: Lucas can believe his Whitely sentence.
- Maudlin, T. 1996. Between the motion and the act... *Psyche* 2:40-51.
- McCall, S. 1999. Can a Turing machine know that the Godel sentence is true? *Journal of Philosophy* 96:525-32.
- McCullough, D. 1996. Can humans escape Godel? *Psyche* 2:57-65.
- McDermott, D. 1996. [Star] Penrose is wrong. *Psyche* 2:66-82.
- Penrose, R. 1989. *The Emperor's New Mind*. Oxford University Press.
We are non-algorithmic as we can see Godel sentences of any algorithm.
- Penrose, R. 1990. *Precis of The Emperor's New Mind*. Behavioral and Brain Sciences 13:643-705.
Much debate over the "non-algorithmic insight" in seeing Godel sentences.
- Penrose, R. 1992. Setting the scene: The claim and the issues. In (D. Broadbent, ed) *The Simulation of Human Intelligence*. Blackwell.
An argument from the halting problem to the nonalgorithmicity of mathematical thought. Addresses objections: that the algorithm is unknowable, unsound, everchanging, environmental, or random. New physical laws may be involved.
- Penrose, R. 1994. *Shadows of the Mind*. Oxford University Press.
- Penrose, R. 1996. Beyond the doubting of a shadow. *Psyche* 2:89-129.
A reply to Chalmers, Feferman, Maudlin, McDermott, etc.
- Priest, G. 1994. Godel's theorem and the mind... again. In (M. Michael & J. O'Leary-Hawthorne, eds) *Philosophy in Mind: The Place of Philosophy in the Study of Mind*. Kluwer.
- Putnam, H. 1985. Reflexive reflections. *Erkenntnis* 22:143-153.
A generalized Godelian argument: if our prescriptive inductive competence is formalizable, then we could not know that such a formalization is correct.
- Robinson, W.S. 1992. Penrose and mathematical ability. *Analysis* 52:80-88.

Penrose's argument depends on our knowledge of the validity of the algorithm we use, and here he equivocates between conscious and unconscious algorithms.

Slezak, P. 1982. Godel's theorem and the mind. *British Journal for the Philosophy of Science* 33:41-52.

General analysis; Lucas commits type/token error; self-ref paradoxes.

Slezak, P. 1983. Descartes's diagonal deduction. *British Journal for the Philosophy of Science* 34:13-36.

Cogito was a diagonal argument; connection to Godel, Lucas, Minsky, Nagel.

Smart, J.J.C. 1961. Godel's theorem, Church's theorem, and mechanism. *Synthese* 13:105-10.

A machine could escape the Godelian argument by inductively ascertaining its own syntax. With comments on the relevance of ingenuity.

Tymoczko, T. 1991. Why I am not a Turing Machine: Godel's theorem and the philosophy of mind. In (J. Garfield, ed) *_Foundations of Cognitive Science_*. Paragon House.

Weak defense of Lucas; response to Putnam, Bowie, Dennett.

Wang, H. 1974. *_From Mathematics to Philosophy_*. London.

Webb, J. 1968. Metamathematics and the philosophy of mind. *Philosophy of Science* 35:156-78.

Webb, J. 1980. *_Mechanism, Mentalism and Metamathematics_*. Kluwer.

Whitely, C. 1962. Minds, machines and Godel: A reply to Mr. Lucas. *Philosophy* 37:61-62.

Humans get trapped too: "Lucas cannot consistently assert this formula".

Yu, Q. 1992. Consistency, mechanicalness, and the logic of the mind. *Synthese* 90:145-79.

4.1c The Chinese Room (Searle)

Searle, J.R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-57.

Implementing a program is not sufficient for mentality, as someone could e.g. implement a "Chinese-speaking" program without understanding Chinese. So strong AI is false, and no program is sufficient for consciousness.

Searle, J.R. 1984. *_Minds, Brains and Science_*. Harvard University Press.

Axiomatizes the argument: Syntax isn't sufficient for semantics, programs are syntactic, minds are semantic, so no program is sufficient for mind.

Searle, J.R. 1987. Minds and brains without programs. In (C. Blakemore, ed) *_Mindwaves_*. Blackwell.

More on the arguments against AI, e.g. the Chinese room and considerations about syntax and semantics. Mind is a high-level physical property of brain.

Searle, J.R. 1990. Is the brain's mind a computer program? *Scientific American* 262(1):26-31.

On the status of the Chinese Room argument, ten years on.

Anderson, D. 1987. Is the Chinese room the real thing? *Philosophy* 62:389-93.

Boden, M. 1988. Escaping from the Chinese Room. In *_Computer Models of Mind_*. Cambridge University Press.

A procedural account of how computers might have understanding and semantics.

Ben-Yami, H. 1993. A note on the Chinese room. *Synthese* 95:169-72.

A fully functional Chinese room is impossible, as it (for instance) could not say what the time is.

Bynum, T.W. 1985. Artificial intelligence, biology, and intentional states. *Metaphilosophy* 16:355-77.

A chess-playing machine embodied as a robot could have intentional states. Reference requires input/output, computation, and context.

Cam, P. 1990. Searle on strong AI. *Australasian Journal of Philosophy* 68:103-8.

Criticizes Searle's "conclusion" that brains are needed for intentionality, notes that even a homunculus has intentional states. A misinterpretation.

Carleton, L. 1984. Programs, language understanding, and Searle. *Synthese* 59:219-30.

Arguing against Searle on a number of fronts, somewhat unconvincingly.

Chalmers, D.J. 1992. Subsymbolic computation and the Chinese Room. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

Gives an account of symbolic vs. subsymbolic computation, and argues that the latter is less vulnerable to the Chinese-room intuition, as representations there are not computational tokens.

Churchland, P.M. & Churchland, P.S. 1990. Could a machine think? *Scientific American* 262(1):32-37.

Artificial mentality is possible, not through classical AI but through brain-like AI. Argues the syntax/semantics point using an analogy with electromagnetism and luminance.

Cohen, L.J. 1986. What sorts of machines can understand the symbols they use? *Aristotelian Society Supplement* 60:81-96.

Cole, D.J. 1984. Thought and thought experiments. *Philosophical Studies* 45:431-44.

Lots of thought experiments like Searle's, against Searle. Searle's argument is like Leibniz's "mill" argument, with similar level confusions. Nice but patchy.

Cole, D.J. 1991. Artificial intelligence and personal identity. *Synthese* 88:399-417.

In the Chinese room, neither the person nor the system understands: a virtual person does. This person isn't the system, just as a normal person isn't a body. Follows from the "Kornese" room, which has two distinct understanders.

Cole, D.J. 1991. Artificial minds: Cam on Searle. *Australasian Journal of Philosophy* 69:329-33.

Cole, D.J. 1994. The causal powers of CPUs. In (E. Dietrich, ed) *Thinking Computers and Virtual Persons*. Academic Press.

Copeland, B.J. 1993. The curious case of the Chinese gym. *Synthese* 95:173-86. Advocates the systems reply, and criticizes Searle's "Chinese Gym" response to connectionism: Searle (like those he accuses) confuses a simulation with the thing being simulated. Nice.

Dennett, D.C. 1987. Fast thinking. In *The Intentional Stance*. MIT Press. Argues with Searle on many points. A little weak.

Double, R. 1983. Searle, programs and functionalism. *Nature and System* 5:107-14.

The homunculus doesn't have access to the system's intentionality. The

syntax/semantics relation is like the neurophysiology/mind relation.

Dyer, M. 1990. Intentionality and computationalism: minds, machines, Searle and Harnad. *Journal of Experimental and Theoretical Artificial Intelligence* 2:303-19.

Reply to Searle/Harnad: systems reply, level confusions, etc.

Dyer, M. 1990. Finding lost minds. *Journal of Experimental and Theoretical Artificial Intelligence* 2:329-39.

Reply to Harnad 1990: symbols, other minds, physically embodied algorithms.

Fields, C. 1984. Double on Searle's Chinese Room. *Nature and System* 6:51-54.

Double's argument implies that the brain isn't the basis of intentionality.

Fisher, J. 1988. The wrong stuff: Chinese rooms and the nature of understanding. *Philosophical Investigations* 11:279-99.

Fodor, J.A. 1991. Yin and Yang in the Chinese Room. In (D. Rosenthal, ed) *The Nature of Mind*. Oxford University Press.

The Chinese room isn't even implementing a Turing machine, because it doesn't use proximal causation. With a reply by Searle.

Globus, G. 1991. Deconstructing the Chinese room. *Journal of Mind and Behavior* 12:377-91.

Gozzano, S. 1995. Consciousness and understanding in the Chinese room. *Informatica* 19:653-56.

Hanna, P. 1985. Causal powers and cognition. *Mind* 94:53-63.

Argues that Searle is confused, and underestimates computers. Weak.

Harnad, S. 1989. Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence* 1:5-25.

Non-symbolic function is necessary for mentality. Trying hard to work out a theory of why the Chinese Room shows what it does. Nice but wrong.

Harnad, S. 1990. Lost in the hermeneutical hall of mirrors. *Journal of Experimental and Theoretical Artificial Intelligence* 2:321-27.

Reply to Dyer 1990: on the differences between real and as-if intentionality.

Hauser, L. 1997. Searle's Chinese box: Debunking the Chinese room argument. *Minds and Machines* 7:199-226.

Hayes, P., Harnad, S., Perlis, D. & Block, N. 1992. Virtual symposium on virtual mind. *Minds and Machines* 2.

A discussion about the Chinese room, symbol grounding, and so on.

Hofstadter, D.R. 1981. Reflections on Searle. In (D. Hofstadter & D. Dennett, eds) *The Mind's I*, pp. 373-382. Basic Books.

Searle is committing a level confusion, and understates the complexity of the case. We can move from the CR to a brain (with a demon) by twiddling knobs, and the systems reply should work equally well in both cases.

Jacquette, D. 1989. Searle's intentionality thesis. *Synthese* 80:267-75.

Searle's view implies that intentional causation is not efficient causation.

Jacquette, D. 1989. Adventures in the Chinese Room. *Philosophy and Phenomenological Research* 49:605-23.

If we had microfunctional correspondence, the CR argument would fail. With points about the status of abstract/biological intentionality. A bit weak.

Searle, J.R. 1989. Reply to Jacquette. *Philosophy and Phenomenological Research* 49:701-8.

Jacquette misses the point of the argument. Also, biological and abstract intentionality are quite compatible.

Jacquette, D. 1990. Fear and loathing (and other intentional states) in Searle's Chinese Room. *Philosophical Psychology* 3:287-304.

Reply to Searle on CR, central control, biological intentionality & dualism.

Jahren, N. 1990. Can semantics be syntactic? *Synthese* 82:309-28.

Against Rapaport's Korean Room argument -- syntax isn't enough.

Korb, K. 1991. Searle's AI program. *Journal of Experimental and Theoretical Artificial Intelligence* 3:283-96.

The Chinese room doesn't succeed as an argument about semantics. At best it might succeed as an argument about consciousness.

Maloney, J.C. 1987. The right stuff. *Synthese* 70:349-72.

Defends Searle against all kinds of objections.

Melnyk, A. 1996. Searle's abstract argument against strong AI. *Synthese* 108:391-419.

Moor, J.H. 1988. The pseudorealization fallacy and the Chinese Room argument. In (J. Fetzer, ed) *_Aspects of AI_*. D. Reidel.

Computational systems must also meet performance criteria.

Newton, N. 1989. Machine understanding and the Chinese Room. *Philosophical Psychology* 2:207-15.

A program can possess intentionality, even if not consciousness.

Obermeier, K.K. 1983. Wittgenstein on language and artificial intelligence: The Chinese-room thought-experiment revisited. *Synthese* 56:339-50.

Pfeifer, K. 1992. Searle, strong AI, and two ways of sorting cucumbers. *Journal of Philosophical Research* 17:347-50.

Rapaport, W. 1984. Searle's experiments with thought. *Philosophy of Science* 53:271-9.

Comments on Cole, and some general material on syntax and semantics.

Rey, G. 1986. What's really going on in Searle's 'Chinese Room'. *Philosophical Studies* 50:169-85.

Recommends the systems reply, and a causal account of semantics. Discusses the relevance of wide and narrow notions of content, and the tension between Searle's positive and negative proposals.

Roberts, L. 1990. Searle's extension of the Chinese Room to connectionist machines. *Journal of Experimental and Theoretical Artificial Intelligence* 2:185-7.

In arguing against the relevance of the serial/parallel distinction to mental states, Searle becomes a formalist. A nice point.

Russow, L-M. 1984. Unlocking the Chinese Room. *Nature and System* 6:221-8.

Searle's presence in the room destroys the integrity of the system, so that it is no longer a proper implementation of the program.

Seidel, A. 1988. Searle on the biological basis of cognition. *Analysis* 48:26-28.

Seidel, A. 1989. Chinese Rooms A, B and C. *Pacific Philosophical Quarterly* 20:167-73.

A person running the program, with interpretations at hand, would understand. Point-missing.

Sharvy, R. 1985. Searle on programs and intentionality. Canadian Journal of Philosophy Supplement 11:39-54.

Argues against Searle, but misses the point for the most part.

Sloman, A. 1986. Did Searle attack Strong Strong AI or Weak Strong AI? In (Cohn & Thomas, eds) Artificial Intelligence and its Applications. Chichester.

Suits, D. 1989. Out of the Chinese Room. Computing and Philosophy Newsletter 4:1-7.

Story about homunculi within homunculi. Fun.

Teng, N.Y. 2000. A cognitive analysis of the Chinese room argument. Philosophical Psychology 13:313-24.

Thagard, P. 1986. The emergence of meaning: An escape from Searle's Chinese Room. Behaviorism 14:139-46.

Get semantics computationally via induction and functional roles.

Weiss, T. 1990. Closing the Chinese room. Ratio 3:165-81.

Searle-in-the-room isn't in a position to know about the system's first-person states. Intrinsic intentionality is an incoherent notion.

Whitmer, J.M. 1983. Intentionality, artificial intelligence, and the causal powers of the brain. Auslegung 10:194-210.

Defending Searle's position, with remarks on the "causal powers" argument.

4.1d Machine Consciousness, Misc [see also 1.8b]

Angel, L. 1989. How to Build a Conscious Machine. Westview Press.

Angel, L. 1994. Am I a computer? In (E. Dietrich, ed) Thinking Computers and Virtual Persons. Academic Press.

Arrington, R. 1999. Machines, consciousness, and thought. Idealistic Studies 29:231-243.

Barnes, E. 1991. The causal history of computational activity: Maudlin and Olympia. Journal of Philosophy 88:304-16.

Response to Maudlin 1989. True computation needs active, not passive causation, so Maudlin's machine isn't really computing.

Birnbacher, D. 1995. Artificial consciousness. In (T. Metzinger, ed) Conscious Experience. Ferdinand Schoningh.

Bringsjord, S. 1992. What Robots Can and Can't Be. Kluwer.

Bringsjord, S. 1994. Could, how could we tell if, and should -- androids have inner lives? In (K.M. Ford, C. Glymour, & P. Hayes, eds) Android Epistemology. MIT Press.

Caplain, G. 1995. Is consciousness a computational property? Informatica 19:615-19.

Coles, L.S. 1993. Engineering machine consciousness. AI Expert 8:34-41.

D'Aquili, E.G. & Newberg, A.B. 1996. Consciousness and the machine. Zygon 31:235-52.

Danto, A. 1960. On consciousness in machines. In (S. Hook, ed) Dimensions of Mind. New York University Press.

Dennett, D.C. 1994. The practical requirements for making a conscious robot.

Philosophical Transactions of the Royal Society A 349:133-46.

Dennett, D.C. 1995. Cog: Steps toward consciousness in robots. In (T. Metzinger, ed) _Conscious Experience_. Ferdinand Schoningh.

Glennan, S.S. 1995. Computationalism and the problem of other minds. *Philosophical Psychology* 8:375-88.

Gunderson, K. 1968. Robots, consciousness and programmed behaviour. *British Journal for the Philosophy of Science* 19:109-22.

Gunderson, K. 1969. Cybernetics and mind-body problems. *Inquiry* 12:406-19.

Gunderson, K. 1971. _Mentality and Machines_. Doubleday.

Hillis, D. 1998. Can a machine be conscious? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness II_. MIT Press.

Kirk, R. 1986. Sentience, causation and some robots. *Australasian Journal of Philosophy* 64:308-21.

One could model brain states with monadic states and appropriate connections.
But surely that's not intelligent -- the causation has the wrong form. Nice.

Kitamura, T., Tahara, T., & Asami, K. 2000. How can a robot have consciousness? *Advanced Robotics* 14:263-275.

Lucas, J.R. 1994. A view of one's own (conscious machines). *Philosophical Transactions of the Royal Society, Series A* 349:147-52.

Maudlin, T. 1989. Computation and consciousness. *Journal of Philosophy* 86:407-32.

Computational state is not sufficient for consciousness, as it can be instantiated by a mostly inert object. A nice thought-experiment, raising questions about the relevance of counterfactuals to consciousness.

McCarthy, J. 1996. Making robots conscious of their mental states. In (S. Muggleton, ed) _Machine Intelligence 15_. Oxford University Press.

McGinn, C. 1987. Could a machine be conscious? In (C. Blakemore & S. Greenfield, ed) _Mindwaves_. Blackwell. Reprinted in _The Problem of Consciousness_ (Blackwell, 1980).

Of course, as we are machines. But what *sort* of machines are conscious, and in virtue of what properties? Remarks on artefacts, life, functionalism, and computationalism. So far, we don't know what makes the brain conscious.

Puccetti, R. 1967. On thinking machines and feeling machines. *British Journal for the Philosophy of Science* 18:39-51.

Machines can think but can't feel, so aren't persons.

Putnam, H. 1964. Robots: machines or artificially created life? *Journal of Philosophy* 61:668-91. Reprinted in _Mind, Language, and Reality_ (Cambridge University Press, 1975).

Various arguments and counter-arguments re machine consciousness and civil liberties. Problems of machine consciousness are analogous to problems of human consciousness. The structural basis of the two may well be the same.

Putnam, H. 1967. The mental life of some machines. In (H. Castaneda, ed) _Intentionality, Minds and Perception_. Wayne State University Press. Reprinted in _Mind, Language, and Reality_ (Cambridge University Press, 1975).

More on TMs: explaining their psychology via preference functions.

Schlagel, R. 1999. Why not artificial consciousness or thought? *Minds and Machines* 9:3-28.

- Scriven, M. 1953. The mechanical concept of mind. *Mind*.
To speak of a conscious machine is to commit a semantic mistake.
Consciousness presupposes life and non-mechanism. Later retracted.
- Stubenbergh, L. 1992. What is it like to be Oscar? *Synthese* 90:1-26.
Argues that AI systems like Pollock's Oscar needn't be conscious. Blindsight tells us that complex perceptual processing can go on unconsciously.
- Thompson, D. 1965. Can a machine be conscious? *British Journal for the Philosophy of Science* 16:36.
Accepting machine consciousness would have few philosophical consequences, whereas rejecting it would tend to commit one to epiphenomenalism.
- van de Vete, D. 1971. The problem of robot consciousness. *Philosophy and Phenomenological Research* 32:149-65.
- Ziff, P. 1959. The feelings of robots. *Analysis*.
Of course robots can't think: they're not alive, so this gives us good reason not to rely on behavior. With replies by J.J.C. Smart, N. Smart.
- 4.1e Machine Thought, Misc

- Bringsjord, S. 1998. Cognition is not computation: The argument from irreversibility. *Synthese*.
- Burks, A.W. 1973. Logic, computers, and men. *Proceedings and Addresses of the American Philosophical Association* 46:39-57.
Arguing that a finite deterministic automaton can perform all natural human functions. With remarks on the logical organization of computers.
- Cohen, L.J. 1955. Can there be artificial minds? *Analysis* 16:36-41.
Subservience to known or knowable rules is incompatible with mentality.
- Copeland, B. J. 2000. Narrow versus wide mechanism: Including a re-examination of Turing's views on the mind-machine issue. *Journal of Philosophy* 97:5-33.
- Dennett, D.C. 1985. Can machines think? In *_How We Know_* (Shafto).
Defends the Turing Test, among other things.
- Dretske, F. 1985. Machines and the mental. *Proceedings and Addresses of the American Philosophical Association* 59:23-33.
- Dretske, F. 1993. Can intelligence be artificial? *Philosophical Studies* 71:201-16.
Intelligence requires not just action or thought, but the governance of action by thought, which requires a history. "Wired-up" systems lack the explanatory connection between thought and action, so are not intelligent.
- Dreyfus, H.L. 1972. *_What Computers Can't Do_*. Harper and Row.
Computers follow rules, people don't.
- Hauser, L. 1993. Why isn't my pocket calculator a thinking thing? *Minds and Machines* 3:3-10.
- Henley, T.B. 1990. Natural problems and artificial intelligence. *Behavior and Philosophy* 18:43-55.
On the philosophical importance of criteria for intelligence. With remarks on Searle, the Turing test, attitudes to AI, and ethical considerations.
- Kearns, J.T. 1997. Thinking machines: Some fundamental confusions. *Minds and Machines* 7:269-87.

Lanier, J. 1998. Three objections to the idea of artificial intelligence. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Mackay, D.M. 1951. Mind-life behavior in artifacts. *British Journal for the Philosophy of Science* 2:105-21.

Mackay, D.M. 1952. Mentality in machines. *Aristotelian Society Supplement* 26:61-86.

Manning, R.C. 1987. Why Sherlock Holmes can't be replaced by an expert system. *Philosophical Studies* 51:19-28.

An expert system would lack Holmes' ability to raise the right questions, sort out relevant data, and determine what data are in need of explanation.

Mays, W. 1952. Can machines think? *Philosophy* 27:148-62.

McCarthy, J. 1979. Ascribing mental qualities to machines. In (M. Ringle, ed) *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

Negley, G. 1951. Cybernetics and theories of mind. *Journal of Philosophy* 48:574-82.

Preston, B. 1995. The ontological argument against the mind-machine hypothesis. *Philosophical Studies* 80:131-57.

Lucas, Searle, and Penrose all fall prey to "dual-description" fallacies.

Puccetti, R. 1966. Can humans think? *Analysis*.

Rapaport, W. 1993. Because mere calculating isn't thinking: Comments on Hauser's "Why isn't my pocket calculator a thinking thing?". *Minds and machines* 3:11-20.

Scriven, M. 1960. The compleat robot: A prolegomena to androidology. In (S. Hook, ed) *Dimensions of Mind*. New York University Press.

A machine could possess every characteristic of human thought: e.g. freedom, creativity, learning, understanding, perceiving, feeling.

Spilsbury, R.J. 1952. Mentality in machines. *Aristotelian Society Supplement* 26:27-60.

4.2 Computation and Representation

4.2a Symbols and Symbol Systems [see also 2.1a, 4.3e]

Cummins, R. 1996. Why there is no symbol grounding problem? In *Representations, Targets, and Attitudes*. MIT Press.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42:335-346.

AI symbols are empty and meaningless. They need to be "grounded" in something, e.g. sensory projection. Maybe connectionism can do the trick?

Harnad, S. 1992. Connecting object to symbol in modeling cognition. In (A. Clark & R. Lutz, eds) *Connectionism in Context*. Springer-Verlag.

On the limitations of symbol systems, and the potential for grounding symbols in sensory icons and categorical perception, e.g. with neural networks.

Kosslyn, S.M. & Hatfield, G. 1984. Representation without symbol systems. *Social Research* 51:1019-1045.

MacDorman, K.F. 1997. How to ground symbols adaptively. In (S. O'Neillain, P. McKevitt, & E. MacAogain, eds) Two Sciences of Mind. John Benjamins.

Newell, A. 1980. Physical symbol systems. *Cognitive Science* 4:135-83.

Newell, A. & Simon, H.A. 1981. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery* 19:113-26. Reprinted in (J. Haugeland, ed) Mind Design. MIT Press.
On computer science, AI, & the Physical Symbol System Hypothesis.

Robinson, W.S. 1995. Brain symbols and computationalist explanation. *Minds and Machines* 5:25-44.

Sun, R. 2000. Symbol grounding: a new look at an old idea. *Philosophical Psychology* 13:149-172.

4.2b Computational Semantics

Fodor, J.A. 1978. Tom Swift and his procedural grandmother. *Cognition* 6:229-47. Reprinted in RePresentations (MIT Press, 1980).
Against procedural semantics; it's a rerun of verificationism.

Hadley, R.F. 1990. Truth conditions and procedural semantics. In (P. Hanson, ed) Information, Language and Cognition. University of British Columbia Press.

Johnson-Laird, P. 1977. Procedural semantics. *Cognition* 5:189-214.

Johnson-Laird, P. 1978. What's wrong with Grandma's guide to procedural semantics: A reply to Jerry Fodor. *Cognition* 9:249-61.

McDermott, D. 1978. Tarskian semantics, or no notation without denotation. *Cognitive Science* 2:277-82.
On the virtues of denotational semantics for AI. Notation without denotation, as found in many AI systems, leads to castles in the air.

Perlis, D. 1991. Putting one's foot in one's head -- Part 1: Why. *Nous* 25:435-55.

Perlis, D. 1994. Putting one's foot in one's head -- Part 2: How. In (E. Dietrich, ed) Thinking Computers and Virtual Persons. Academic Press.

Rapaport, W.J. 1988. Syntactic semantics: Foundations of computational natural language understanding. In (J. Fetzer, ed) Aspects of AI. Kluwer.

Rapaport, W.J. 1995. Understanding understanding: Syntactic semantics and computational cognition. *Philosophical Perspectives* 9:49-88.

Smith, B. 1988. On the semantics of clocks. In (J. Fetzer, ed) Aspects of AI. Kluwer.

Smith, B. 1987. The correspondence continuum. *CSLI-87-71*.

Wilks, Y. 1982. Some thoughts on procedural semantics. In (W. Lehnert, ed) Strategies for Natural Language Processing. Lawrence Erlbaum.

Wilks, Y. 1990. Form and content in semantics. *Synthese* 82:329-51.
Criticism of McDermott's views on semantics, logic and natural language.

Winograd, T. 1985. Moving the semantic fulcrum. *Linguistics and Philosophy* 8:91-104.

Woods, W. 1981. Procedural semantics as a theory of meaning. In (A. Joshi,

B. Weber, & I. Sag) *_Elements of Discourse Understanding_*. Cambridge University Press.

Woods, W. 1986. Problems in procedural semantics. In (Z. Pylyshyn & W. Demopolous, eds) *_Meaning and Cognitive Structure_*. Ablex.
With commentaries by Haugeland, J.D. Fodor.

4.2c Implicit/Explicit Rules and Representations

Clark, A. 1991. In defense of explicit rules. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.
Argues that we need explicit rules for flexibility, adaptability, and representational redescription. With remarks on eliminativism.

Cummins, R. 1986. Inexplicit information. In (M. Brand & R. Harnish, eds) *_The Representation of Knowledge and Belief_*. University of Arizona Press.
On various kinds of representation of knowledge or belief without explicit tokens: control-implicit, domain-implicit, and procedural information.
The key distinction is representation vs. execution of a rule.

Davies, M. 1995. Two notions of implicit rules. *Philosophical Perspectives* 9:153-83.

Hadley, R.F. 1990. Connectionism, rule-following, and symbolic manipulation. *Proc AAAI*.
Some rules are learnt so quickly that representation must be explicit.

Hadley, R.F. 1993. Connectionism, explicit rules, and symbolic manipulation. *Minds and Machines* 3.

Hadley, R.F. 1995. The 'explicit-implicit' distinction. *Minds and Machines* 5:219-42.

Kirsh, D. 1990. When is information explicitly represented? In (P. Hanson, ed) *_Information, Language and Cognition_*. University of British Columbia Press.

Skokowski, P.G. 1994. Can computers carry content "inexplicitly"? *Minds and Machines* 4:333-44.
Cummins' account of inexplicit information fails, as even "executed" rules must be represented in the system. With remarks on the Chinese room.

4.2d AI without Representation?

Bechtel, W. 1996. Yet another revolution? Defusing the dynamical system theorists' attack on mental representations. Manuscript.

Brooks, R. 1991. Intelligence without representation. *Artificial Intelligence* 47:139-159.
We don't need explicit representation; the world can do the job instead.
Use embodied, complete systems, starting simple and working incrementally.

Clark, A. and Toribio, J. 1994. Doing without representing. *Synthese* 101:401-31.
A discussion of anti-representationalism in situated robotics and the dynamic systems movement (Brooks, Beer, van Gelder). These arguments appeal to overly simple domains, and a modest notion of representation survives.

Keijzer, F.A. 1998. Doing without representations which specify what to do. *Philosophical Psychology* 11:269-302.

Kirsh, D. 1991. Today the earwig, tomorrow man? *Artificial Intelligence* 47:161-184.

van Gelder, T. 1995. What might cognition be if not computation? *Journal of Philosophy* 92:345-81.
Argues for a dynamic-systems conception of the mind that is non-computational and non-representational. Uses an analogy with the Watt steam governor to argue for a new kind of dynamic explanation.

4.2e Miscellaneous

Chrisley, R.L. 1994. Taking embodiment seriously: Nonconceptual content and robotics. In (K.M. Ford, C. Glymour, & P. Hayes, eds) *_Android Epistemology_*. MIT Press.

Dietrich, E. 1988. Computers, intentionality, and the new dualism. Manuscript.

Dreyfus, H.L. 1979. A framework for misrepresenting knowledge. In (M. Ringle, ed) *_Philosophical Perspectives in Artificial Intelligence_*. Humanities Press.
On the problems with context-free symbolic representation.

Fields, C. 1994. Real machines and virtual intentionality: An experimentalist takes on the problem of representational content. In (E. Dietrich, ed) *_Thinking Computers and Virtual Persons_*. Academic Press.

Haugeland, J. 1981. Semantic engines: An introduction to mind design. In (J. Haugeland, ed) *_Mind Design_*. MIT Press.

Robinson, W.S. 1995. Direct representation. *Philosophical Studies* 80:305-22.
On Searle's critique of computational explanation, contrasted with Gallistel's use thereof. The real issue is computation on indirect vs. direct representations; direct computationalism is an attractive view.

4.3 Philosophy of Connectionism

4.3a Connectionism and Compositionality (Fodor/Pylyshyn)

Fodor, J.A. & Pylyshyn, Z.W. 1988. Connectionism and cognitive architecture. *Cognition* 28:3-71.

Connectionist models can't explain cognitive systematicity and productivity, as their representations lack compositional structure. The allures of connectionism are illusory; it's best used as an implementation strategy.

Aizawa, K. 1997. Explaining systematicity. *Mind and Language* 12:115-36.

Aizawa, K. 1997. The role of the systematicity argument in classicism and connectionism. In (S. O'Nuallain, ed) *_Two Sciences of Mind_*. John Benjamins.

Aizawa, K. 1997. Exhibiting verses explaining systematicity: A reply to Hadley and Hayward. *Minds and Machines* 7:39-55.

Antony, M. 1991. Fodor and Pylyshyn on connectionism. *Minds and Machines* 1:321-41.

Fodor and Pylyshyn's argument is an invalid instance of inference to the best explanation, as there is much to explain than systematicity. Connectionism and classicism may be compatible even without implementation, in any case.

Aydede, M. 1997. Language of thought: The connectionist contribution. *Minds and Machines* 7:57-101.

Butler, K. 1991. Towards a connectionist cognitive architecture. *Mind and Language* 6:252-72.

Connectionism can make do with unstructured representations, as long have they have the right causal relations between them.

Butler, K. 1993. Connectionism, classical cognitivism, and the relation between cognitive and implementational levels of analysis. *Philosophical Psychology* 6:321-33.

Contra Chalmers 1993, F&P's argument doesn't apply at the implementational level. Contra Chater and Oaksford 1990, connectionism can't be purely implementational, but some implementational details can be relevant.

Butler, K. 1993. On Clark on systematicity and connectionism. *British Journal for the Philosophy of Science* 44:37-44.

Argues against Clark on holism and the conceptual truth of systematicity.

Butler, K. 1995. Compositionality in cognitive models: The real issue. *Philosophical Studies* 78:153-62.

Chalmers, D.J. 1990. Syntactic transformations on distributed representations. *Connection Science* 2:53-62.

An experimental demonstration that connectionist models can handle structure-sensitive operations in a non-classical way, transforming structured representations of active sentences to passive sentences.

Chalmers, D.J. 1993. Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology* 6:305-319.

Points out a structural flaw in F&P's argument, and traces the problem to a lack of appreciation of distributed representation. With some empirical results on structure sensitive processing, and some remarks on explanation.

Chater, N. & Oaksford, M. 1990. Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition* 34:93-107.

Implementation can make a difference at the algorithmic level.

Christiansen, M.H. & Chater, N. 1994. Generalization and connectionist language learning. *Mind and Language* 9:273-87.

Cummins, R. 1996. Systematicity. *Journal of Philosophy* 93:591-614.

Fetzer, J.H. 1992. Connectionism and cognition: Why Fodor and Pylyshyn are wrong. In (A. Clark & R. Lutz, eds) *_Connectionism in Context_*. Springer-Verlag.

Fodor, J.A. & McLaughlin, B.P. 1990. Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition* 35:183-205.

Smolensky's weak compositionality is useless; and tensor product architecture can't support systematicity, as nonexistent tokens can't play a causal role.

Fodor, J.A. 1997. Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition* 62:109-19.

Garcia-Carpintero, M. 1996. Two spurious varieties of compositionality. *Minds and Machines* 6:159-72.

Garfield, J. 1997. Mentalese not spoken here: Computation, cognition, and causation. *Philosophical Psychology* 10:413-35.

Guarini, M. 1996. Tensor products and split-level architecture: Foundational issues in the classicism-connectionism debate. *Philosophy of Science* 63:S239-47.

Hadley, R.F. 1997. Cognition, systematicity, and nomic necessity. *Mind and Language* 12:137-53.

Hadley, R.F. 1994. Systematicity in connectionist language learning. *Mind and Language* 9:247-72.

Argues that existing connectionist models do not achieve an adequate systematicity in learning; they fail to generalize to handle structures with novel constituents.

Hadley, R.F. 1994. Systematicity revisited. *Mind and Language* 9:431-44.

Hadley, R.F. & Hayward, M.B. 1997. Strong semantic systematicity from Hebbian connectionist learning. *Minds and Machines* 7:1-55.

Hadley, R.F. 1997. Cognition, systematicity, and nomic necessity. *Mind and Language* 12:137-53.

Hadley, R.F. 1997. Explaining systematicity: A reply to Kenneth Aizawa. *Minds and Machines* 12:571-79.

Hawthorne, J. 1989. On the compatibility of connectionist and classical models. *Philosophical Psychology* 2:5-16.

Localist connectionist models may not be able to handle structured presentation, but appropriate distributed models can.

Horgan, T. & Tienson, J. 1991. Structured representations in connectionist systems? In (Davis, ed) *_Connectionism: Theory and Practice_*.

A discussion of how connectionism might achieve "effective syntax" without implementing a classical system.

Matthews, R.J. 1994. Three-concept monte: Explanation, implementation, and systematicity. *Synthese* 101:347-63.

F&P deal a sucker bet: on their terms, connectionism could never give a non-implementational explanation of systematicity, as the notions are construed in a manner specific to classical architectures.

Matthews, R.J. 1997. Can connectionists explain systematicity? *Mind and Language* 12:154-77.

McLaughlin, B.P. 1992. Systematicity, conceptual truth, and evolution. In *_Philosophy and the Cognitive Sciences_*.

Against responses to Fodor and Pylyshyn claiming that cognitive theories needn't explain systematicity. Contra Clark, the conceptual truth of systematicity won't help. Contra others, nor will evolution.

McLaughlin, B.P. 1993. The connectionism/classicism battle to win souls. *Philosophical Studies* 71.

Argues that no connectionist model so far has come close to explaining systematicity. Considers the models of Elman, Chalmers, and Smolensky.

Niklasson, L.F. & van Gelder, T. 1994. On being systematically connectionist. *Mind and Language* 9:288-302.

Pollack, J.B. 1990. Recursive distributed representations. *Artificial Intelligence* 46:77-105.

Develops a connectionist architecture -- recursive auto-associative memory -- that can recursively represent compositional structures in distributed form.

Rowlands, M. 1994. Connectionism and the language of thought. *British Journal for the Philosophy of Science* 45:485-503.

F&P's argument confuses constituent structure with logical/sentential structure. Connectionism is a psychotechnic project, whereas propositional description is a psychosemantic project.

Schroder, J. 1998. Knowledge of rules, causal systematicity, and the language of thought. *Synthese* 117:313-330.

Smolensky, P. 1987. The constituent structure of connectionist mental states. *Southern Journal of Philosophy Supplement* 26:137-60.

F&P ignore distributed representation and interaction effects.

Smolensky, P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159-216.

Develops a connectionist architecture that represents compositional structures as tensor products of distributed representations.

Smolensky, P. 1991. Connectionism, constituency and the language of thought. In (B. Loewer & G. Rey, eds) *Meaning in Mind: Fodor and his Critics*. Blackwell.

Connectionism can do compositionality its own way, including both weak compositionality (with context effects) or strong compositionality (via tensor products).

Smolensky, P. 1995. Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In (C. Macdonald, ed) *Connectionism: Debates on Psychological Explanation*. Blackwell.

van Gelder, T. 1990. Compositionality: A connectionist variation on a classical theme. *Cognitive Science* 14:355-84.

Connectionism can do compositionality functionally. All one needs is the right functional relation between representations; physical concatenation is not necessary.

van Gelder, T. 1991. Classical questions, radical answers. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

On connectionism as a Kuhnian paradigm shift in cognitive science, with emphasis on the implications of functional compositionality and distributed representations.

4.3b Representation in Connectionism

Butler, K. 1995. Representation and computation in a deflationary assessment of connectionist cognitive science. *Synthese* 104:71-97.

Clark, A. 1989. Connectionism, nonconceptual content, and representational redescription. Manuscript.

On some troubles connectionism has with higher-order knowledge. Contrasts Cussins, Karmiloff-Smith on development. Subsymbols without symbols are blind.

Clark, A. 1993. *Associative Engines: Connectionism, Concepts, and Representational Change*. MIT Press.

Clark, A. & Karmiloff-Smith, A. 1994. The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language* 8:487-519.

On the importance of representational redescription, and on the limits of connectionist networks in cross-domain knowledge transfer. What does a true believer need, above behavior: conceptual combination, real-world fluency?

Cummins, R. 1991. The role of representation in connectionist explanation of cognitive capacities. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Connectionism isn't really radical. There's no new concept of representation

or of learning, and cognition can still be the manipulation of semantically structured representations.

Cussins, A. 1990. The connectionist construction of concepts. In (M. Boden, ed) *The Philosophy of AI*. Oxford University Press.

Connectionism builds up concepts from the nonconceptual level. From nonconceptual content (e.g. perceptual experiences) to the emergence of objectivity.

Garzon, F. 2000. A connectionist defence of the inscrutability thesis. *Mind and Language* 15:465-480.

Garzón, F. 2000. State space semantics and conceptual similarity: reply to Churchland. *Philosophical Psychology* 13:77-96.

Goschke, T. & Koppelberg, D. 1990. Connectionism and the semantic content of internal representation. *Review of International Philosophy* 44:87-103.

Goschke, T. & Koppelberg, D. 1991. The concept of representation and the representation of concepts in connectionist models. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.
On correlational semantics and context-dependent representations.

Hatfield, G. 1991. Representation and rule-instantiation in connectionist systems. In (T. Horgan & J. Tienson, eds) *Connectionism and the Philosophy of Mind*. Kluwer.

Some remarks on psychology & physiology. Even connectionism uses psychological concepts.

Hatfield, G. 1991. Representation in perception and cognition: Connectionist affordances. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Haybron, D.M. 2000. The causal and explanatory role of information stored in connectionist networks. *Minds and Machines* 10:361-380.

Laakso, A. & Cottrell, G. 2000. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical Psychology* 13:47-76.

Place, U.T. 1989. Toward a connectionist version of the causal theory of reference. *Acta Analytica* 4:71-97.

Ramsey, W. 1995. Rethinking distributed representation. *Acta Analytica* 10:9-25.

Ramsey, W. 1997. Do connectionist representations earn their explanatory keep? *Mind and Language* 12:34-66.

Argues that talk of representations has no explanatory role in connectionist theory, and can be discarded. It can't be understood along the lines of the teleo-informational or classical frameworks.

Schopman, J. & Shawky, A. 1996. Remarks on the impact of connectionism on our thinking about concepts. In (P. Millican & A. Clark, eds) *Machines and Thought*. Oxford University Press.

Tye, M. 1987. Representation in pictorialism and connectionism. *Southern Journal of Philosophy Supplement* 26:163-184.

Pictorialism isn't compatible with language of thought, but connectionism might be.

van Gelder, T. 1991. What is the D in PDP? In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Argues that distributed representation is best analyzed in terms of superposition of representation, not in terms of extendedness.

4.3c Connectionism and Eliminativism

Ramsey, W., Stich, S.P. & Garon, J. 1991. Connectionism, eliminativism and the future of folk psychology. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) Philosophy and Connectionist Theory. Lawrence Erlbaum.

Connectionism implies eliminativism, as connectionist systems do not have functionally discrete contentful states, and folk psychology is committed to functional discreteness of propositional attitudes.

Bickle, J. 1993. Connectionism, eliminativism, and the semantic view of theories. *Erkenntnis*.

Outlines the semantic view of scientific theories, and applies it to the connectionism/eliminativism debate. There's no reason why folk psychology shouldn't be reducible, in a homogeneous or heterogeneous way.

Botterill, G. 1994. Beliefs, functionally discrete states, and connectionist networks. *British Journal for the Philosophy of Science* 45:899-906.

Distinguishes active from dispositional beliefs: the former are realized discretely in activation patterns, the latter nondiscretely in weights, which is all that folk psychology needs.

Clapin, H. 1991. Connectionism isn't magic. *Minds and Machines* 1:167-84.

Commentary on Ramsey/Stich/Garon. Connectionism has symbols that interact, and has propositional modularity in processing if not in storage.

Clark, A. 1989. Beyond eliminativism. *Mind and Language* 4:251-79.

Connectionism needn't imply eliminativism, as higher levels may have a causal role, if not causal completeness. Also, it may not tell the whole story.

Clark, A. 1990. Connectionist minds. *Proceedings of the Aristotelian Society* 90:83-102.

Responding to eliminativist challenge via cluster analysis and recurrence.

Davies, M. 1991. Concepts, connectionism, and the language of thought. (W. Ramsey, S. Stich, & D. Rumelhart, eds) Philosophy and Connectionist Theory. Lawrence Erlbaum.

Argues that our conception of thought requires causal systematicity, which requires a language of thought. Connectionist systems are not causally systematic, so connectionism leads to eliminativism.

Egan, F. 1995. Folk psychology and cognitive architecture. *Philosophy of Science* 62:179-96.

Forster, M. & Saidel, E. 1994. Connectionism and the fate of folk psychology. *Philosophical Psychology* 7:437-52.

Contra Ramsey, Stich, and Garon, connectionist representations can be seen to be functionally discrete on an appropriate analysis of causal relevance.

Horgan, T., and Tienson, J. 1995. Connectionism and the commitments of folk psychology. *Philosophical Perspectives* 9:127-52.

O'Brien, G. 1991. Is connectionism commonsense? *Philosophical Psychology* 4:165-78.

O'Leary-Hawthorne, J. 1994. On the threat of eliminativism. *Philosophical Studies* 74:325-46.

A dispositional construal of beliefs and desires can distinguish the relevant active states (via counterfactuals) and is compatible with FP, so internals can't threaten FP. With remarks on Davidson, overdetermination, etc.

Place, U.T. 1992. Eliminative connectionism: Its implications for a return to an empiricist/behaviorist linguistics. *Behavior and Philosophy* 20:21-35.

Ramsey, W. 1994. Distributed representation and causal modularity: A rejoinder to Forster and Saidel. *Philosophical Psychology* 7:453-61.

Upon examination, the model of Forster and Saidel 1994 does not exhibit features that are both distributed and causally discrete.

Smolensky, P. 1995. On the projectable predicates of connectionist psychology: A case for belief. In (C. Macdonald, ed) *_Connectionism: Debates on Psychological Explanation_*. Blackwell.

Stich, S. & Warfield, T. 1995. Reply to Clark and Smolensky: Do connectionist minds have beliefs? In (C. Macdonald, ed) *_Connectionism: Debates on Psychological Explanation_*. Blackwell.

4.3d The Connectionist/Classical Debate

Adams, F., Aizawa, K. & Fuller, G. 1992. Rules in programming languages and networks. In (J. Dinsmore, ed) *_The Symbolic and Connectionist Paradigms: Closing the Gap_*. Lawrence Erlbaum.

The distinction between programming languages and networks is neutral on rule-following, etc, so there's nothing really new about connectionism.

Aizawa, K. 1994. Representations without rules, connectionism, and the syntactic argument. *Synthese* 101:465-92.

Bringsjord, S. 1991. Is the connectionist-logician debate one of AI's wonderful red herrings? *Journal of Theoretical and Experimental Artificial Intelligence* 3:319-49.

A detailed analysis purporting to show that connectionism and "logicism" are compatible, as Turing machines can do everything a neural network can. Entertaining, but misunderstands subsymbolic processing.

Broadbent, D. 1985. A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General* 114:189-92.

Distributed models are at the implementational, not computational, level.

Chandrasekaran, B., Goel, A. & Allemang, D. 1988. Connectionism and information-processing abstractions. *AI Magazine* 24-34.

Connectionism won't affect AI too much, as AI is concerned with the information-processing (task) level. With greater modularity, connectionism will look more like traditional AI.

Corbi, J.E. 1993. Classical and connectionist models: Levels of description. *Synthese* 95:141-68.

Dawson, M.R.W., Medler, D.A., & Berkeley, I.S.N. 1997. PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology* 10:25-40.

Dennett, D.C. 1986. The logical geography of computational approaches: A view from the east pole. In (M. Brand & R. Harnish, eds) *_The Representation of Knowledge and Belief_*. University of Arizona Press.

Drawing the battle-lines: High Church Computationalism at the "East Pole", New Connectionism, Zen Holism, etc, at various locations on the "West Coast". With remarks on connectionism, and on AI as thought-experimentation.

Dennett, D.C. 1991. Mother Nature versus the walking encyclopedia. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.

Reiterating the value of connectionism, especially biological plausibility.

Dinsmore, J. (ed) 1992. _The Symbolic and Connectionist Paradigms: Closing the Gap_. Lawrence Erlbaum.

Dyer, M. 1991. Connectionism versus symbolism in high-level cognition. In (T. Horgan & J. Tienson, eds) _Connectionism and the Philosophy of Mind_. Kluwer.

Garson, J.W. 1991. What connectionists cannot do: The threat to Classical AI. In (T. Horgan & J. Tienson, eds) _Connectionism and the Philosophy of Mind_. Kluwer.

Connectionism and classicism aren't necessarily incompatible on symbolic discreteness, causal role, functional discreteness, constituency, representation of rules.

Garson, J.W. 1994. No representations without rules: The prospects for a compromise between paradigms in cognitive science. *Mind and Language* 9:25-37.

Garson, J.W. 1994. Cognition without classical architecture. *Synthese* 100:291-306.

Horgan, T. & Tienson, J. 1987. Settling into a new paradigm. *Southern Journal of Philosophy Supplement* 26:97-113.

On connectionism, basketball, and representation without rules. Responses to the "syntactic" and "semantic" arguments against connectionism. Nice.

Horgan, T. & Tienson, J. 1989. Representation without rules. *Philosophical Perspectives* 17:147-74.

Cognition uses structured representations without high-level rules, and connectionism is better at accounting for this. With remarks on exceptions to psychological laws, and the crisis in traditional AI.

Horgan, T. & Tienson, J. 1994. Representations don't need rules: Reply to James Garson. *Mind and Language* 9:1-24.

McClelland, J.L. & Rumelhart, D.E. 1985. Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General* 114:193-7.

Response to Broadbent 1985: Distributed models are at the algorithmic level. Elucidating the low-level/high-level relation via various analogies.

McLaughlin, B.P. & Warfield, F. 1994. The allure of connectionism reexamined. *Synthese* 101:365-400.

Argues that symbolic systems such as decision trees are as good at learning and pattern recognition as connectionist networks, and it is just as plausible that they are implemented in the brain.

Rey, G. 1991. An explanatory budget for connectionism and eliminativism. In (T. Horgan & J. Tienson, eds) _Connectionism and the Philosophy of Mind_. Kluwer.

Challenges connectionism to explain things that the classical approach seems to handle better: the structure, systematicity, causal role, and grain of propositional attitudes, their rational relations, and conceptual stability.

4.3e Subsymbolic Computation (Smolensky)

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1-23.

Connectionism offers a complete account at the subsymbolic level, rather than an approximate account at the symbolic level.

Berkeley, I. 2000. What the #*\$%! is a subsymbol? *Minds and machines* 10:1-14.

Chalmers, D.J. 1992. Subsymbolic computation and the Chinese Room. In (J. Dinsmore, ed) *The Symbolic and Connectionist Paradigms: Closing the Gap*. Lawrence Erlbaum.

Explicates the distinction between symbolic and subsymbolic computation, and argues that connectionism can better handle the emergence of semantics from syntax, due to the non-atomic nature of its representations.

Clark, A. 1993. Superpositional connectionism: A reply to Marinov. *Minds and Machines* 3:271-81.

Hofstadter, D.R. 1983. Artificial intelligence: Subcognition as computation. In (F. Machlup, ed) *The Study of Information: Interdisciplinary Messages*. Wiley. Reprinted as "Waking up from the Boolean dream" in *Metamagical Themas*. Basic Books.

AI needs statistical emergence. For real semantics, symbols must be decomposable, complex, autonomous -- i.e. active.

Marinov, M. 1993. On the spuriousness of the symbolic/subsymbolic distinction. *Minds and Machines* 3:253-70.

Argues with Smolensky: symbolic systems such as decision trees have all the positive features of neural networks (flexibility, lack of brittleness), and can represent concepts as sets of subconcepts. With a reply by Clark.

Rosenberg, J. 1990. Treating connectionism properly: Reflections on Smolensky. *Psychological Research* 52:163.

Rejects Smolensky's PTC, as the proper interaction of the microscopic and macroscopic levels would take a "miracle".

Smolensky, P. 1987. Connectionist AI, symbolic AI, and the brain. *AI Review* 1:95-109.

On connectionist networks as subsymbolic dynamic systems.

4.3f Philosophy of Connectionism, Misc.

Bechtel, W. 1985. Are the new PDP models of cognition cognitivist or associationist? *Behaviorism* 13:53-61.

Bechtel, W. 1986. What happens to accounts of mind-brain relations if we forgo an architecture of rules and representations? *Philosophy of Science Association* 1986, 159-71.

On the relationship between connectionism, symbol processing, psychology and neuroscience.

Bechtel, W. 1987. Connectionism and the philosophy of mind. *Southern Journal of Philosophy Supplement* 26:17-41. Reprinted in (W. Lycan, ed) *Mind and Cognition* (Blackwell, 1990).

Lots of questions about connectionism.

Bechtel, W. 1988. Connectionism and rules and representation systems: Are they compatible? *Philosophical Psychology* 1:5-16.

There's room for both styles within a single mind. The rule-based level needn't be autonomous; the connectionist level plays a role in pattern recognition, concepts, and so on.

Bechtel, W. & Abrahamson, A. 1990. Beyond the exclusively propositional era. *Synthese* 82:223-53.

An account of the shift from propositions to pattern recognition in the study of cognition: knowing-how, imagery, categorization, connectionism.

Bechtel, W. & Abrahamsen, A.A. 1992. Connectionism and the future of folk psychology. In (R. Burton, ed) *Minds: Natural and Artificial*. SUNY Press.

Bechtel, W. 1993. The case for connectionism. *Philosophical Studies* 71:119-54.

Bickle, J. 1995. Connectionism, reduction, and multiple realizability. *Behavior and Philosophy* 23:29-39.

Bradshaw, D.E. 1991. Connectionism and the specter of representationalism. In (T. Horgan & J. Tienson, eds) *_Connectionism and the Philosophy of Mind_*. Kluwer.

Argues that connectionism allows for a more plausible epistemology of perception, compatible with direct realism rather than representationalism. With remarks on Fodor and Pylyshyn's argument against Gibson.

Churchland, P.M. 1989. On the nature of theories: A neurocomputational perspective. *Minnesota Studies in the Philosophy of Science* 14. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).

Connectionism will revolutionize our review of scientific theories: >From the deductive-nomological view to descent in weight-space. Some cute analogies.

Churchland, P.M. 1989. On the nature of explanation: A PDP approach. In *_A Neurocomputational Perspective_*. MIT Press.

We achieve explanatory understanding not through the manipulation of propositions but through the activation of prototypes.

Churchland, P.S. & Sejnowski, T. 1989. Neural representation and neural computation. In (L. Nadel, ed) *_Neural Connections, Mental Computations_*. MIT Press.

Implications of connectionism and neuroscience for our concept of mind.

Clark, A. 1989. *_Microcognition_*. MIT Press.

All kinds of stuff on connectionism and philosophy.

Clark, A. 1990. Connectionism, competence and explanation. *British Journal for the Philosophy of Science* 41:195-222.

Connectionism separates processing from competence. Instead of hopping down Marr's levels (theory->process), connectionism goes (1) task (2) low-level performance (3) extract theory from process. Cute.

Cummins, R. & Schwarz, G. 1987. Radical connectionism. *Southern Journal of Philosophy Supplement* 26:43-61.

On computation and representation in AI and connectionism, and on problems for radical connectionism in reconciling these without denying representation or embracing mystery.

Cummins, R. & Schwarz, G. 1991. Connectionism, computation, and cognition. In (T. Horgan & J. Tienson, eds) *_Connectionism and the Philosophy of Mind_*. Kluwer.

Explicates computationalism, and discusses ways in which connectionism might end up non-computational: if causal states cross-classify representational states, or if transitions between representations aren't computable.

Cummins, R. 1995. Connectionist and the rationale constraint on cognitive explanations. *Philosophical Perspectives* 9:105-25.

Davies, M. 1989. Connectionism, modularity and tacit knowledge. *British Journal for the Philosophy of Science* 40:541-55.

Argues that connectionist networks don't have tacit knowledge of modular theories (as representations lack the appropriate structure, etc.).

Globus, G.G. 1992. Derrida and connectionism: Differance in neural nets. *Philosophical Psychology* 5:183-97.

Hatfield, G. 1990. Gibsonian representations and connectionist symbol-processing: prospects for unification. *Psychological Research* 52:243-52.

Gibson is compatible with connectionism. In both, we can have rule-instantiation without rule-following.

Horgan, T. & Tienson, J. (eds) 1991. *_Connectionism and the Philosophy of Mind_*. Kluwer.

Horgan, T. & Tienson, J. 1996. *_Connectionism and the Philosophy of Psychology_*. MIT Press.

Horgan, T. 1997. Connectionism and the philosophical foundations of cognitive science. *Metaphilosophy* 28:1-30.

Humphreys, G.W. 1986. Information-processing systems which embody computational rules: The connectionist approach. *Mind and Language* 1:201-12.

Legg, C.R. 1988. Connectionism and physiological psychology: A marriage made in heaven? *Philosophical Psychology* 1:263-78.

Litch, M. 1997. Computation, connectionism and modelling the mind. *Philosophical Psychology* 10:357-364.

Lloyd, D. 1989. Parallel distributed processing and cognition: Only connect? In *_Simple Minds_*. MIT Press.

An overview: local/distributed/featural representations; explanation in connectionism (how to avoid big mush); relation to neuroscience; explicit representations of rules vs weight matrices.

Lycan, W.G. 1991. Homuncular functionalism meets PDP. In (W. Ramsey, S. Stich, & D. Rumelhart, eds) *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.

On various ways in which connectionism relates to representational homuncular functionalism, e.g. on implementation, eliminativism, and explanation.

Macdonald, C. 1995. *_Connectionism: Debates on Psychological Explanation_*. Blackwell.

Ramsey, W. & Stich, S.P. 1990. Connectionism and three levels of nativism. *Synthese* 82:177-205.

How connectionism bears on the nativism debate. Conclusion: not too much.

Ramsey, W., Stich, S.P. & Rumelhart, D.M. (eds) 1991. *_Philosophy and Connectionist Theory_*. Lawrence Erlbaum.

Rosenberg, J. 1989. Connectionism and cognition. Bielefeld Report. Criticism of Churchland's connectionist epistemology.

Sehon, S. 1998. Connectionism and the causal theory of action explanation. *Philosophical Psychology* 11:511-532.

Shanon, B. 1992. Are connectionist models cognitive? *Philosophical Psychology*.

In some senses of "cognitive", yes; in other senses, no. Phenomenological, theoretical, and sociological perspectives. Toward meaning-laden models.

Sterelny, K. 1990. Connectionism. In *_The Representational Theory of Mind_*. Blackwell.

Waskan, J. & Bechtel, W. 1997. Directions in connectionist research: Tractable computations without syntactically structured representations. *Metaphilosophy* 28:31-62.

4.3g Foundational Empirical Issues

Clark, A. 1994. Representational trajectories in connectionist learning. *Minds and Machines* 4:317-32.

On how to get connectionist networks to learn about structured task domains. Concentrates on incremental learning, and other developmental/scaffolding strategies. With remarks on systematicity.

Clark, A. & Thornton, S. 1997. Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences* 20:57-66.

Cliff, D. 1990. Computational neuroethology: A provisional manifesto. Manuscript.

Criticizes connectionism for not being sufficiently rooted in neuroscience, and for not being grounded in the world.

Dawson, M.R.W. & Schopflocher, D.P. 1992. Autonomous processing in parallel distributed processing networks. *Philosophical Psychology* 5:199-219.

Hanson, S. & Burr, D. 1990. What connectionist models learn. *Behavioral and Brain Sciences*.

What's new to connectionism is not learning or representation but the way that learning and representation interact.

Kaplan, S., Weaver, M. & French, R.M. 1990. Active symbols and internal models: Towards a cognitive connectionism. *AI and Society*.

Addresses behaviorist/associationist charges. Connectionism needs recurrent circuits to support active symbols.

Kirsh, D. 1987. Putting a price on cognition. *Southern Journal of Philosophy Supplement* 26:119-35.

On the importance of variable binding, and why it's hard with connectionism.

Lachter, J. & Bever, T. 1988. The relation between linguistic structure and associative theories of language learning. *Cognition* 28:195-247.

Criticism of connectionist language models. They build in too much.

Mills, S. 1989. Connectionism, the classical theory of cognition, and the hundred step constraint. *Acta Analytica* 4:5-38.

Nelson, R. 1989. Philosophical issues in Edelman's neural darwinism. *Journal of Experimental and Theoretical Artificial Intelligence* 1:195-208.

On the relationship between ND, PDP and AI. All are computational.

Oaksford, M., Chater, N. & Stenning, K. 1990. Connectionism, classical cognitive science and experimental psychology. *AI and Society*.

Connectionism is better at explaining empirical findings about mind.

Pinker, S. & Prince, A. 1988. On language and connectionism. *Cognition* 28:73-193.

Extremely thorough criticism of the R&M past-tense-learning model, with arguments on why connectionism can't handle linguistic rules.

4.4 Dynamical Systems

Bechtel, W. 1996. Yet another revolution? Defusing the dynamical system theorists' attack on mental representations. Manuscript.

Clark, A. 1998. Time and mind. *Journal of Philosophy* 95:354-76.

Eliasmith, C. 1996. The third contender: A critical examination of the dynamicist theory of cognition. *Philosophical Psychology* 9:441-63.

Eliasmith, C. 1997. Computation and dynamical models of mind. *Minds and Machines* 7:531-41.

Foss, J.E. 1992. Introduction to the epistemology of the brain: Indeterminacy, micro-specificity, chaos, and openness. *Topoi* 11:45-57.

On the brain as a vector-processing system, and the problems raised by indeterminacy, chaos, and so on. With morals for cognitive science.

Freeman, W. 1997. Nonlinear neurodynamics of intentionality. *Journal of Mind and Behavior* 18:291-304.

Garson, J.W. 1995. Chaos and free will. *Philosophical Psychology* 8:365-74.

Garson, J.W. 1996. Cognition poised at the edge of chaos: A complex alternative to a symbolic mind. *Philosophical Psychology* 9:301-22.

Garson, J.W. 1997. Syntax in a dynamic brain. *Synthese* 110:343-355.

Garson, J.W. 1998. Chaotic emergence and the language of thought. *Philosophical Psychology* 11:303-315.

Giunti, M. 1995. Dynamic models of cognition. In (T. van Gelder & R. Port, eds) *_Mind as Motion_*. MIT Press.

Giunti, M. 1996. *_Computers, Dynamical Systems, and the Mind_*. Oxford University Press.

Globus, G. 1992. Toward a noncomputational cognitive science. *Journal of Cognitive Neuroscience* 4:299-310.

Hooker, C.A. & Christensen, W.D. 1998. Towards a new science of the mind: Wide content and the metaphysics of organizational properties in nonlinear dynamic models. *Mind and Language* 13:98-109.

Horgan, T. & Tienson, J. 1992. Cognitive systems as dynamic systems. *Topoi* 11:27-43.

Horgan, T. & Tienson, J. 1994. A nonclassical framework for cognitive science. *Synthese* 101:305-45.

Keijzer, F.A. & Bem, S. 1996. Behavioral systems interpreted as autonomous agents and as coupled dynamical systems: A criticism. *Philosophical Psychology* 9:323-46.

Sloman, A. 1993. The mind as a control system. In (C. Hookway & D. Peterson, eds) *_Philosophy and Cognitive Science_*. Cambridge University Press.

van Gelder, T. & Port, R. 1995. *_Mind as Motion: Explorations in the Dynamics of Cognition_*. MIT Press.

van Gelder, T. 1995. What might cognition be if not computation? *Journal of Philosophy* 92:345-81.

Argues for a dynamic-systems conception of the mind that is non-computational and non-representational. Uses an analogy with the Watt steam governor to argue for a new kind of dynamic explanation.

van Gelder, T. 1997. Connectionism, dynamics, and the philosophy of mind. In (M. Carrier & P. Machamer, eds) *_Mindsapes: Philosophy, Science, and the Mind_*. Pittsburgh University Press.

van Gelder, T. 1998. The dynamical hypothesis in cognitive science.

Behavioral and Brain Sciences 21:615-28.

4.5 Foundational Questions in AI

4.5a The Nature of AI

Buchanan, B. 1988. AI as an experimental science. In (J. Fetzer, ed) *_Aspects of AI_*. D. Reidel.

Bundy, A. 1990. What kind of field is AI? In (D. Partridge & Y. Wilks, eds) *_The Foundations of Artificial Intelligence: A Sourcebook_*. Cambridge University Press.

Dennett, D.C. 1978. AI as philosophy and as psychology. In (M. Ringle, ed) *_Philosophical Perspectives on Artificial Intelligence_*. Humanities Press. Reprinted in *_Brainstorms_* (MIT Press, 1978).

AI as detailed armchair psychology and as thought-experimental epistemology. Implications for mind: e.g. a solution to the problem of homuncular regress.

Glymour, C. 1988. AI is philosophy. In (J. Fetzer, ed) *_Aspects of AI_*. D. Reidel.

Kukla, A. 1989. Is AI an empirical science? *Analysis* 49:56-60.

No, AI is an a priori science that uses empirical methods; its status is similar to that of mathematics.

Kukla, A. 1994. Medium AI and experimental science. *Philosophical Psychology* 7:493-5012.

On the status of "medium AI", the study of intelligence in computational systems (not just humans). Contra to many, this is not an empirical science, but a combination of (experimental) mathematics and engineering.

Nakashima, H. 1999. AI as complex information processing. *Minds and Machines* 9:57-80.

4.5b Levels of Analysis (Marr, etc)

Bechtel, W. 1994. Levels of description and explanation in cognitive science. *Minds and Machines* 4:1-25.

Cleeremans, A. & French, R.M. 1996. From chicken squawking to cognition: Levels of description and the computational approach in psychology. *Psychologica Belgica* 36:5-29.

Foster, C. 1990. *_Algorithms, abstraction and implementation_*. Academic Press.

Outlines a theory of the equivalence of algorithms.

Horgan, T. & Tienson, J. 1992. Levels of description in nonclassical cognitive science. *Philosophy* 34, Supplement.

Generalizes Marr's levels to: cognitive state-transitions, mathematical state-transitions, implementation. Discusses these with respect to connectionism, dynamical systems, and computation below the cognitive level.

Houng, Y. 1990. Classicism, connectionism and the concept of level. Dissertation, Indiana University.

On levels of organization vs. levels of analysis.

Marr, D. 1982. *_Vision_*. Freeman.

Defines computational, algorithmic and implementational levels.

McClamrock, R. 1990. Marr's three levels: a re-evaluation. *Minds and Machines* 1:185-196.

On different kinds of level-shifts: organizational and contextual changes.
There are more than three levels available.

Newell, A. 1982. The knowledge level. *Artificial Intelligence* 18:81-132.

Newell, A. 1986. The symbol level and the knowledge level. In (Z. Pylyshyn & W. Demopolous, eds) *Meaning and Cognitive Structure*. Ablex.

With commentaries by Smith, Dennett.

Peacocke, C. 1986. Explanation in computational psychology: Language, perception and level 1.5. *Mind and Language* 1:101-23.

Psychological explanation is typically somewhere *between* the computational and algorithmic levels.

Sticklen, J. 1989. Problem-solving architectures at the knowledge level. *Journal of Experimental and Theoretical Artificial Intelligence* 1:233-247.

4.5c The Frame Problem

Dennett, D.C. 1984. Cognitive wheels: The frame problem of AI. In (Hookaway, ed) *Minds, Machines and Evolution*. Cambridge University Press.

General overview.

Dreyfus, H.L. & Dreyfus, S. 1987. How to stop worrying about the frame problem even though it's computationally insoluble. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

FP is an artifact of computational explicitness. Contrast human commonsense know-how, with relevance built in. Comparison to expert/novice distinction.

Fetzer, J.H. 1990. The frame problem: Artificial intelligence meets David Hume. *International Journal of Expert Systems* 3:219-232.

Fodor, J.A. 1987. Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

FP is Hamlet's problem: when to stop thinking. It's equivalent to the general problem of non-demonstrative inference.

Haugeland, J. 1987. An overview of the frame problem. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

The FP may be a consequence of the explicit/implicit rep distinction. Use "complicit" reps instead, and changes will be carried along intrinsically.

Hayes, P. 1987. What the frame problem is and isn't. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

FP is a relatively narrow problem, Some, e.g. Fodor, wrongly equate FP with the "Generalized AI Problem".

Janlert, L. 1987. Modeling change: The frame problem. In (Z. Pylyshyn, ed) *The Robot's Dilemma*. Ablex.

Korb, K. 1998. The frame problem: An AI fairy tale. *Minds and Machines* 8:317-351.

Lormand, E. 1990. Framing the frame problem. *Synthese* 82:353-74.

Criticizes Dennett's, Haugeland's and Fodor's construals of the FP.

Maloney, J.C. 1988. In praise of narrow minds. In (J. Fetzer, ed) *Aspects of AI*. D. Reidel.

McCarthy, J. & Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. In (Meltzer & Michie, eds) Machine Intelligence 4. Edinburgh University Press.

McDermott, D. 1987. We've been framed: Or, Why AI is innocent of the frame problem. In (Z. Pylyshyn, ed) The Robot's Dilemma. Ablex.
Solve frame problem by using the sleeping-dog strategy -- keeping things fixed unless there's a reason to suppose otherwise.

Pollock, J.L. 1997. Reasoning about change and persistence: A solution to the frame problem. *Nous* 31:143-169.

Pylyshyn, Z.W. (ed) 1987. The Robot's Dilemma. Ablex.
Lots of papers on the frame problem.

4.5d AI Methodology

Birnbaum, L. 1991. Rigor mortis: A response to Nilsson's 'Logic and artificial intelligence'. *Artificial Intelligence* 47:57-78.

Chalmers, D.J., French, R.M. & Hofstadter, D.R. 1992. High-level perception, representation, and analogy: A critique of AI methodology. *Journal of Experimental and Theoretical Artificial Intelligence*.
AI must integrate perception and cognition in the interest of flexible representation. Current models ignore perception and the development of representation, but this cannot be separated from later cognitive processes.

Clark, A. 1986. A biological metaphor. *Mind and Language* 1:45-64.
AI should look at biology.

Clark, A. 1987. The kludge in the machine. *Mind and Language* 2:277-300.

Dascal, M. 1992. Why does language matter to artificial intelligence? *Minds and Machines* 2:145-174.

Dreyfus, H.L. 1981. From micro-worlds to knowledge: AI at an impasse. In (J. Haugeland, ed) Mind Design. MIT Press.
Micro-worlds don't test true understanding, and frames and scripts are doomed to leave out too much. Explicit representation can't capture intelligence.

Dreyfus, H.L. & Dreyfus, S.E. 1988. Making a mind versus modeling the brain: AI at a crossroads. *Daedalus*.
History of AI (boo) and connectionism (qualified hooray). And Husserl/Heidegger/Wittgenstein. Quite nice.

Hadley, R.F. 1991. The many uses of 'belief' in AI. *Minds and Machines* 1:55-74.
Various AI approaches to belief: syntactic, propositional/meaning-based, information, tractability, discoverability, and degree of confidence.

Haugeland, J. 1979. Understanding natural language. *Journal of Philosophy* 76:619-32. Reprinted in (W. Lycan, ed) Mind and Cognition (Blackwell, 1990).
AI will need holism: interpretational, common-sense, situational, existential.

Kirsh, D. 1991. Foundations of AI: The big issues. *Artificial Intelligence* 47:3-30.
Identifying the dividing lines: pre-eminence of knowledge, embodiment, language-like kinematics, role of learning, uniformity of architecture.

Marr, D. 1977. Artificial intelligence: A personal view. *Artificial Intelligence* 9:37-48.

AI usually comes up with Type 2 (algorithmic) theories, when Type 1 (info processing) theories might be more useful -- at least if they exist.

McDermott, D. 1981. Artificial intelligence meets natural stupidity. In (J. Haugeland, ed) *Mind Design*. MIT Press.
Problems in AI methodology: wishful mnemonics, oversimplifying natural language concepts, and never implementing programs. Entertaining.

McDermott, D. 1987. A critique of pure reason. *Computational Intelligence* 3:151-60.
Criticism of logicism (i.e. reliance on deduction) in AI.

Nilsson, N. 1991. Logic and artificial intelligence. *Artificial Intelligence* 47:31-56.

Partridge, D. & Wilks, Y. (eds) 1990. *The Foundations of Artificial Intelligence: A Sourcebook*. Cambridge University Press.
Lots of papers on various aspects of AI methodology. Quite thorough.

Preston, B. 1993. Heidegger and artificial intelligence. *Philosophy and Phenomenological Research* 53:43-69.
On the non-represented background to everyday activity, and environmental interaction in cognition. Criticizes cognitivism, connectionism, looks at Agre/Chapman/Brooks, ethology, anthropology for support.

Pylyshyn, Z.W. 1979. Complexity and the study of artificial and human intelligence. In (M. Ringle, ed) *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

Ringle, M. (ed) 1979. *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.
10 papers on philosophy of AI, psychology and knowledge representation.

Robinson, W.S. 1991. Rationalism, expertise, and the Dreyfuses' critique of AI research. *Southern Journal of Philosophy* 29:271-90.
Defending limited rationalism: i.e. a theory of intelligence below the conceptual level but above the neuronal level.

4.6 Computationalism in Cognitive Science [see also 2.2e] -----

Antony, L. 1997. Feeling fine about the mind. *Philosophy and Phenomenological Research* 57:381-87.

Bickhard, M. 1996. Troubles with computationalism. In (W. O'Donahue & R. Kitchener, eds) *The Philosophy of Psychology*. Sage Publications.

Block, N. 1990. The computer model of mind. In (D. Osherson & E. Smith, eds) *An Invitation to Cognitive Science*, Vol. 3. MIT Press.
Overview of computationalism. Relationship to intentionality, LOT, etc.

Boden, M. 1984. What is computational psychology? *Proceedings of the Aristotelian Society* 58:17-35.

Bringsjord, S. 1994. Computation, among other things, is beneath us. *Minds and Machines* 4:469-88.

Bringsjord, S. & Zenzen, M. 1997. Cognition is not computation: The argument from irreversibility. *Synthese* 113:285-320.

Buller, D.J. 1993. Confirmation and the computational paradigm, or, why do you think they call it artificial intelligence? *Minds and Machines* 3:155-81.

Chalmers, D.J. 1994. A computational foundation for the study of cognition. Manuscript.

Argues for theses of computational sufficiency and computational explanation, resting on the fact that computation provides an abstract specification of causal organization. With replies to many anti-computationalist worries.

Clarke, J. 1972. Turing machines and the mind-body problem. *British Journal for the Philosophy of Science* 23:1-12.

Cummins, R. 1977. Programs in the explanation of behavior. *Philosophy of Science* 44:269-87.

Demopoulos, W. 1987. On some fundamental distinctions of computationalism. *Synthese* 70:79-96.

On analog/digital, representational/nonrepresentational, direct/indirect.

Dietrich, E. 1990. Computationalism. *Social Epistemology*.

What computationalism is, as opposed to computerism & cognitivism. Implies: intentionality isn't special, and we don't make decisions. With commentary.

Dietrich, E. 1989. Semantics and the computational paradigm in computational psychology. *Synthese* 79:119-41.

Argues that computational explanation requires the attribution of semantic content. Addresses Stich's arguments against content, and argues that computers are not formal symbol manipulators.

Double, R. 1987. The computational model of the mind and philosophical functionalism. *Behaviorism* 15:131-39.

Dretske, F. 1985. Machines and the mental. *Proceedings and Addresses of the American Philosophical Association* 59:23-33.

Machines can't even add, let alone think, as the symbols they use aren't meaningful to them. They would need real information based on perceptual embodiment, and conceptual capacities, for meaning to play a real role.

Fetzer, J.H. 1994. Mental algorithms: Are minds computational systems? *Pragmatics and Cognition* 21:1-29.

Fodor, J.A. 1978. Computation and reduction. *Minnesota Studies in the Philosophy of Science* 9. Reprinted in *_RePresentations_* (MIT Press, 1980).

Fodor, J. 2000. *_The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology_*. MIT Press.

Garson, J.W. 1993. Mice in mirrored mazes and the mind. *Philosophical Psychology* 6:123-34.

Computationalism is false, as it can't distinguish the ability to solve a maze for the ability to solve its mirror image. An embodied computational taxonomy is needed, rather than software alone.

Harnad, S. 1996. Computation is just interpretable symbol manipulation; Cognition isn't. *Minds and Machines* 4:379-90.

Horst, S. 1996. *_Symbols, Computation, and Intentionality: A Critique of the Computational Theory of Mind_*. University of California Press.

Horst, S. 1999. Symbols and computation: A critique of the computational theory of mind. *Minds and Machines* 9:347-381

Mellor, D.H. 1984. What is computational psychology? II. *Proceedings of the Aristotelian Society* 58:37-53.

Mellor, D.H. 1989. How much of the mind is a computer. In (P. Slezak, ed)

Computers, Brains and Minds. Kluwer.

Only belief is computational: rest of mind is not.

Nelson, R. 1987. Machine models for cognitive science. *Philosophy of Science*
Argues contra Pylyshyn 1984 that finite state automata are good models for
cognitive science: they are semantically interpretable and process symbols.

Pollock, J. 1989. _How to Build a Person: A Prolegomenon_. MIT Press.

Pylyshyn, Z.W. 1980. Computation and cognition: Issues in the foundation of
cognitive science. *Behavioral and Brain Sciences* 3:111-32.

Pylyshyn, Z.W. 1984. _Computation and Cognition_. MIT Press.
A thorough account of the symbolic/computational view of cognition.

Pylyshyn, Z.W. 1978. Computational models and empirical constraints.
Behavioral and Brain Sciences 1:98-128.

Pylyshyn, Z.W. 1989. Computing and cognitive science. In (M. Posner, ed)
Foundations of Cognitive Science. MIT Press.
An overview of the computational view of mind. On symbols, levels, control
structures, levels of correspondence for computational models, and empirical
methods for determining degrees of equivalence.

Shapiro, S.C. 1995. Computationalism. *Minds and Machines* 5:467-87.

Sterelny, K. 1989. Computational functional psychology: problems and
prospects. In (P. Slezak, ed) _Computers, Brains and Minds_. Kluwer.
Various points on pros and cons of computational psychology.

Tibbetts, P. 1996. Residual dualism in computational theories of mind.
Dialectica 50:37-52.

4.7 Computation and Physical Systems

Boyle, C.F. 1994. Computation as an intrinsic property. *Minds and Machines*
4:451-67.

Chalmers, D.J. 1994. On implementing a computation. *Minds and Machines*
4:391-402.
Gives an account of what it is for a physical system to implement a
computation: the causal structure of the system must mirror the formal
structure of the computation. Answers objections by Searle and others.

Chalmers, D.J. 1996. Does a rock implement every finite-state automaton?
Synthese 108:309-33.
Argues that Putnam's "proof" that every ordinary open system implements
every finite automaton is fallacious. It can be patched up, but an
appropriate account of implementation resists these difficulties.

Chrisley, R.L. 1994. Why everything doesn't realize every computation. *Minds*
and *Machines* 4:403-20.

Cleland, C. 1993. Is the Church-Turing thesis true? *Minds and Machines*
3:283-312.
Many physically realized functions can't be computed by Turing machines:
e.g. "mundane procedures" and continuous functions. So the C-T thesis is
false of these, and maybe even of number-theoretic functions.

Cleland, C.E. 1995. Effective procedures and computable functions. *Minds* and
Machines 5:9-23.

Copeland, B.J. 1996. What is computation? *Synthese* 108:335-59.

Endicott, R.P. 1996. Searle, syntax, and observer-relativity. *Canadian Journal of Philosophy* 26:101-22.

Goel, V. 1991. Notationality and the information processing mind. *Minds and Machines* 1:129-166.

Adapts Goodman's notational systems to explicate computational information processing. What is/isn't a physical notational system (e.g. LOT, symbol systems, connectionism) and why. How to reconcile notational/mental content?

Hardcastle, V.G. 1995. Computationalism. *Synthese* 105:303-17.

Pragmatic factors are vital in connecting the theory of computation with empirical theory, and particularly in determining whether a given system counts as performing a given computation.

Horsten, L. 1995. The Church-Turing thesis and effective mundane procedures. *Minds and Machines* 5:1-8.

MacLennan, B. 1994. "Words lie in our way". *Minds and Machines* 4:421-37.

Miscevic, N. 1996. Computationalism and the Kripke-Wittgenstein paradox. *Proceedings of the Aristotelian Society* 96:215-29.

Scheutz, M. 1999. When physical systems realize functions. *Minds and Machines* 9:161-196.

Searle, J.R. 1990. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association* 64:21-37.

Syntax isn't intrinsic to physics, so computational ascriptions are assigned by observer. Syntax has no causal powers. Brain doesn't process information.

Shagrir, O. 1997. Two dogmas of computationalism. *Minds and Machines* 7:321-44.

Stabler, E. 1987. Kripke on functionalism and automata. *Synthese* 70:1-22.

Disputes Kripke's argument that there is no objective way of determining when a system computes a given function, due to infinite domains and unreliability. Stipulating normal background conditions is sufficient.

Suber, P. 1988. What is software? *Journal of Speculative Philosophy* 2:89-119.

4.8 Philosophy of AI, Misc

Bergadano, F. 1993. Machine learning and the foundations of inductive inference. *Minds and Machines* 3:31-51.

Button, G., Coulter, J., Lee, J.R.E. & Sharrock, W. 1995. *_Computers, Minds, and Conduct_*. Polity Press.

Fetzer, J.H. 1990. *_Artificial Intelligence: Its Scope and Limits_*. Kluwer.

Gips, J. 1994. Toward the ethical robot. In (K.M. Ford, C. Glymour, & P. Hayes, eds) *_Android Epistemology_*. MIT Press.

Haugeland, J. (ed) 1981. *_Mind Design_*. MIT Press.
12 papers on the foundations of AI and cognitive science.

Hayes, P.J., Ford, K.M., & Adams-Webber, J.R. 1994. Human reasoning about artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence* 4:247-63. Reprinted in (E. Dietrich, ed) *_Thinking Computers and Virtual Persons_*. Academic Press.

Krellenstein, M. 1987. A reply to 'Parallel computation and the mind-body problem'. Cognitive Science 11:155-7.

Thagard 1986 is wrong: speed and the like make no fundamental difference.

With Thagard's reply: it makes a difference in practice, if not in principle.

Moody, T.C. 1993. _Philosophy and Artificial Intelligence_. Prentice-Hall.

Preston, B. 1991. AI, anthropocentrism, and the evolution of "intelligence.". Minds and Machines 1:259-277.

Robinson, W.S. 1992. _Computers, Minds, and Robots_. Temple University Press.

Russell, S. 1991. Inductive learning by machines. Philosophical Studies 64:37-64.

A nice paper on the relationship between techniques of theory formation from machine learning and philosophical problems of induction and knowledge.

Rychlak, J.F. 1991. _Artificial Intelligence and Human Reason: A Teleological Critique_. Columbia University Press.

Sloman, A. 1978. _The Computer Revolution in Philosophy_. Harvester.

All about how the computer should change the way we think about the mind.

Thagard, P. 1986. Parallel computation and the mind-body problem. Cognitive Science 10:301-18.

Parallelism does make a difference. Some somewhat anti-functional points.

Thagard, P. 1990. Philosophy and machine learning. Canadian Journal of Philosophy 20:261-76.

Thagard, P. 1991. Philosophical and computational models of explanation. Philosophical Studies 64:87-104.

A comparison of philosophical and AI approaches to explanation: deductive, statistical, schematic, analogical, causal, and linguistic.

Winograd, T. & Flores, F. 1987. _Understanding Computers and Cognition_. Addison-Wesley.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 5: Philosophy of Psychology [561]

=====

Contents

5.1 Issues in Cognitive Science [230]

5.1a Nativism (Chomsky, etc) [37]

5.1b Modularity (Fodor, etc) [30]

5.1c Mental Imagery [45]

5.1d Rationality [31]

5.1e Embodiment [31]

5.1f Animal Cognition [56]

5.2 Aspects of Mind [224]

5.2a Pain and Pleasure [40]

5.2b Emotions [46]

5.2c Dreams [21]

5.2d Memory [32]

5.2e Color [85]

5.3 Philosophy of Psychology and Cognitive Science [80]

5.3a Psychological Laws [13]

- 5.3b Psychology and Neuroscience [33]
- 5.3c Explanation in Cognitive Science [17]
- 5.3d Philosophy of Cognitive Science, Misc [17]
- 5.4 Philosophy of Mind (General) [27]

5.1 Issues in Cognitive Science

5.1a Nativism (Chomsky, etc)

Ariew, A. 1996. Innateness and canalization. *Philosophy of Science Supplement* 63:19-27.

Atherton, M. & Schwarz, R. 1974. Linguistic innateness and its evidence. *Journal of Philosophy* 71:6.

Chomsky, N. 1967. Recent contributions to the theory of innate ideas. *Synthese* 17:2-11.

Chomsky, N. 1969. Linguistics and philosophy. In (S. Hook, ed) *_Language and Philosophy_*. New York University Press.

Reply to Putnam 1967: Putnam underestimates complexity of grammar, etc.

Chomsky, N. 1975. On cognitive capacity. In *_Reflections on Language_*. Pantheon Books.

Chomsky, N. 1980. Discussion of Putnam's comments. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

Chomsky, N. & Fodor, J.A. 1980. The inductivist fallacy. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

Churchland, P.S. 1978. Fodor on language learning. *Synthese* 38:149-59.

Cowie, F. 1997. The logical problem of language acquisition. *Synthese* 111:17-51.

Cowie, F. 1998. *_What's Within?_* Oxford University Press.

Cowie, F. 1998. Mad dog nativism. *British Journal for the Philosophy of Science* 49:227-252.

Cummins, D.D. 1996. Evidence for the innateness of deontic reasoning. *Mind and Language* 11:160-90.

De Rosa, R. 2000. On Fodor's claim that classical empiricists and rationalists agree on the innateness of ideas. *Protosociology* 14:240-269.

Harman, G. 1969. Linguistic competence and empiricism. In (S. Hook, ed) *_Language and Philosophy_*. New York University Press.

Fodor, J.A., Bever, T. & Garrett, M. 1974. The specificity of language skills. In *_The Psychology of Language_*. McGraw-Hill.

Fodor, J.A. 1980. Reply to Putnam. In (M. Piattelli-Palmarini, ed) *_Language and Learning: The Debate between Jean Piaget and Noam Chomsky_*. Harvard University Press.

Fodor, J.A. 1981. The present status of the innateness controversy. In *_Representations_*. MIT Press.

Concepts are undefinable, so primitive, so innate (plus gloss).

- Fodor, J.A. 1980. On the impossibility of acquiring 'more powerful' structures. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard University Press.
- Katz, J. 1966. Innate ideas. In *The Philosophy of Language*. Harper & Row. Overview; poverty of stimulus, unobservable features => rationalism.
- Kaye, L.J. 1993. Are most of our concepts innate? *Synthese* 2:187-217.
- Mehler, J. & Fox, R. (eds) 1985. *Neonate Cognition: Beyond the Blooming Buzzing Confusion*. Lawrence Erlbaum.
- Piattelli-Palmarini, M. (ed) 1980. *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*. Harvard University Press.
An excellent collection of papers & responses by Piaget, Chomsky and others.
- Piattelli-Palmarini, M. 1986. The rise of selective theories: A case study and some lessons from immunology. In (W. Demopoulos, ed) *Language Learning and Concept Acquisition*. Ablex.
- Piattelli-Palmarini, M. 1989. Evolution, selection, and cognition: From learning to parameter setting in biology and in the study of language. *Cognition* 31:1-44.
Why learning is selective and not instructive. Biological analogies, linguistic evidence. Dispense with "learning" as a scientific term.
- Pitt, D. 2000. Nativism and the theory of content. *Protosociology* 14:222-239.
- Putnam, H. 1967. The 'Innateness Hypothesis' and explanatory models in linguistics. *Synthese* 17:12-22. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).
Contra nativism: disputes (1) surprising universals (2) explanation of universals (3) ease of learning (4) relevance of IQ-independence.
- Putnam, H. 1980. What is innate and why. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard University Press.
- Putnam, H. 1980. Comments on Chomsky's and Fodor's replies. In (M. Piattelli-Palmarini, ed) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Harvard University Press.
- Ramsey, W. & Stich, S.P. 1990. Connectionism and three levels of nativism. *Synthese* 82:177-205.
Identifies minimal nativism vs anti-empiricism vs rationalism. Considers the relevance of connectionist networks. Some nativist arguments may survive.
- Samet, J. 1986. Troubles with Fodor's nativism. *Midwest Studies in Philosophy* 10:575-594.
Concepts can be acquired without being learned by symbol-manipulation.
- Samet, J. & Flanagan, O.J. 1989. Innate representations. In (S. Silvers, ed) *Rerepresentation*. Kluwer.
- Sampson, G. 1978. Linguistic universals as evidence for empiricism. *Journal of Linguistics*.
Explain universals via Popper/Simon empirical model.
- Samuels, R. 1998. What brains won't tell us about the mind: A critique of the neurobiological argument against representational nativism. *Mind and Language* 13:548-570.

Schwartz, R. 1995. Is mathematical competence innate? *Philosophy of Science* 62:227-40.

Sterelny, K. 1989. Fodor's nativism. *Philosophical Studies* 55:119-41.

Stich, S.P. (ed) 1975. *_Innate Ideas_*. University of California Press.

Stich, S.P. 1979. Between Chomskian rationalism and Popperian empiricism. *British Journal for the Philosophy of Science* 30:329-47.

Can take middle ground. Anti-empiricism doesn't imply rationalism.

5.1b Modularity (Fodor, etc)

Appelbaum, I. 1998. Fodor, modularity, and speech perception. *Philosophical Psychology* 11:317-330.

Arbib, M. 1989. Modularity, schemas and neurons: A critique of Fodor. In (P. Slezak, ed) *_Computers, Brains and Minds_*. Kluwer.
Against Fodor: modules are smaller, interact strongly, not domain-specific.

Bennett, L.J. 1990. Modularity of mind revisited. *British Journal for the Philosophy of Science* 41:429-36.
Remarks on Shanon and Fodor.

Browne, D. 1996. Cognitive versatility. *Minds and Machines* 6:507-23.

Bruner, J. 1957. On perceptual readiness. *Psychological Review* 65:14-21.
Overview of the original studies on top-down effects in perception.

Cam, P. 1988. Modularity, rationality, and higher cognition. *Philosophical Studies* 53:279-94.

Cam, P. 1990. Insularity and the persistence of perceptual illusion. *Analysis* 50:231-5.

Chien, A.J. 1996. Why the mind may not be modular. *Minds and Machines* 6:1-32.

Churchland, P.M. 1979. *_Scientific Realism and the Plasticity of Mind_*. Cambridge University Press.
Our perception is deeply theory-laden, and potentially very plastic.

Churchland, P.M. 1988. Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor. *Philosophy of Science* 55:167-87. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).
Contra Fodor 1984: observation is theory-laden (built-in or not); supported by neurophysiological evidence; perceptual systems have long-term plasticity.

Currie, G. & Sterelny, K. 2000. How to think about the modularity of mind-reading. *Philosophical Quarterly* 50:145-160.

DesAutels, P. 1995. Two types of theories: The impact of Churchland's perceptual plasticity. *Philosophical Psychology* 8:25-33.

Fodor, J.A. 1983. *_The Modularity of Mind_*. MIT Press.
Perception happens in informationally encapsulated, domain-specific modules. Central systems aren't encapsulated, and so may be impossible to understand.

Fodor, J.A. 1985. *Precis of _The modularity of mind_*. *Behavioral and Brain Sciences* 8:1-42. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).
Summary of MOM (with commentary and reply in the BBS printing).

Fodor, J.A. 1986. The modularity of mind. In (Z. Pylyshyn, ed) *_Meaning and*

Cognitive Structure_. Ablex.

Informal discussion of modularity. With commentaries by Fahlman, Caplan.

Fodor, J.A. 1984. Observation reconsidered. *Philosophy of Science* 51:23-43.
Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

Argues for an observation/theory distinction, and against belief affecting perception.

Fodor, J.A. 1988. A reply to Churchland's 'Perceptual plasticity and theoretical neutrality'. *Philosophy of Science* 55:188-98. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

Churchland is up the creek without a paddle.

Fodor, J.A. 1989. Why should the mind be modular? In (A. George, ed) *_Reflections on Chomsky_*. Blackwell. Reprinted in *_A Theory of Content and Other Essays_* (MIT Press, 1990).

Garfield, J. (ed) 1987. *_Modularity in Knowledge Representation and Natural-Language Understanding_*. MIT Press.

A collection of papers on modularity in language and vision.

Gray, R. 2001. Cognitive modules, synaesthesia and the constitution of psychological natural kinds. *Philosophical Psychology* 14:65-82.

Meyering, T.C. 1994. Fodor's modularity: A new name for an old dilemma. *Philosophical Psychology* 7:39-62.

Olsson, E. 1997. Coherence and the modularity of mind. *Australasian Journal of Philosophy* 75:404-11.

Pylyshyn, Z. 1999. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences* 22:341-365.

Rollins, M. 1994. Deep plasticity: The encoding approach to perceptual change. *Philosophy of Science* 61:39-54.

Ross, J. 1990. Against postulating central systems in the mind. *Philosophy of Science* 57:297-312.

Fodor's arguments for unencapsulated central systems are no good; AI is possible after all.

Shanon, B. 1988. Remarks on the modularity of mind. *British Journal for the Philosophy of Science* 39:331-52.

Criticism of Fodor. Modularity is dynamic, and can be central.

Fodor, J.A. & Pylyshyn, Z.W. 1981. How direct is visual perception?: Some reflections on Gibson's 'ecological approach'. *Cognition* 9:139-96.

'Direct perception' can't correspond to anything. Perception is inferential.

Turvey, M.T., Shaw, R.E., Reed, E.S., & Mace, W.M. 1981. Ecological laws of perceiving and acting: In Reply to Fodor and Pylyshyn. *Cognition* 9:237-304.

Ullman, S. 1980. Against direct perception. *Behavioral and Brain Sciences* 3:333-81.

Vaina, L.M. 1990. What and where in the human visual system: Two hierarchies of visual modules. *Synthese* 83:49-91.

5.1c Mental Imagery [see also 6.2j]

Anderson, J.R. 1978. Arguments concerning representations for mental imagery.

Psychological Review 85.

Audi, R. 1978. The ontological status of mental images. *Inquiry* 21:348-61.

Blachowicz, J. 1997. Analog representation beyond mental imagery. *Journal of Philosophy* 94:55-84.

Block, N. (ed) 1981. *_Imagery_*. MIT Press.

Block, N. 1983. Mental pictures and cognitive science. *Philosophical Review* 93:499-542. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).

Block, N. 1983. The photographic fallacy in the debate about mental imagery. *Nous* 17:651-62.

Brown, R. & Herrnstein, R. 1981. Icons and images. In (N. Block, ed) *_Imagery_*. MIT Press.

Cam, P. 1987. Propositions about images. *Philosophy and Phenomenological Research* 48:335-8.

Candlish, S. 1975. Mental images and pictorial properties. *Mind* 84:260-2.

Chambers, D. & Reisberg, D. 1985. Can mental images be ambiguous?' *Journal of Experimental Psychology: Human Perception and Performance* 11:317-28.

Chambers, D. & Reisberg, D. 1992. What an image depicts depends on what an image means. *Cognitive Psychology* 24:145-74.

Dennett, D.C. 1978. Two approaches to mental images. In *_Brainstorms_*. MIT Press.

Dennett, D.C. 1968. The nature of images and the introspective trap. In *_Content and Consciousness_*. Routledge and Kegan Paul. Reprinted in (N. Block, ed) *_Readings in the Philosophy of Psychology_* (MIT Press, 1980).

Farah, M.J. 1988. Is visual imagery really visual: Some overlooked evidence from neuropsychology. *Psychological Review* 95:307-17.

Finke, R.A. 1989. *_Principles of Mental Imagery_*. MIT Press.

Fodor, J.A. 1975. Imagistic representation. In *_The Language of Thought_*. Harvard University Press.

Glasgow, J.I. 1993. The imagery debate revisited: A computational perspective. *Computational Intelligence* 9:310-33.

Hannay, A. 1971. *_Mental Images: A Defense_*. Allen & Unwin.

Hannay, A. 1973. To see a mental image. *Mind* 82:161-262.

Kind, A. 2001. Putting the image back in imagination. *Philosophy and Phenomenological Research* 62:85-110.

Kosslyn, S.M. & Pomerantz, J. 1977. Imagery, propositions and the form of internal representations. *Cognitive Psychology* 9:52-76. Reprinted in (N. Block, ed) *_Readings in the Philosophy of Psychology_* (MIT Press, 1980).

Kosslyn, S.M. 1981. The medium and the message in mental imagery: A theory. In (N. Block, ed) *_Imagery_*. MIT Press.

Kosslyn, S.M., Pinker, S., Schwartz, S. & Smith, G. 1979. On the demystification of mental imagery. *Behavioral and Brain Sciences* 2:535-81.

- Kosslyn, S.M. 1980. *_Image and Mind_*. Harvard University Press.
- Kosslyn, S.M. 1994. *_Image and Brain: The Resolution of the Imagery Debate_*. MIT Press.
- Maloney, J.C. 1984. Mental images and cognitive theory. *American Philosophical Quarterly* 21:237-47.
- Morris, P.E. & Hampson, P.J. 1983. *_Imagery and Consciousness_*. Academic Press.
- Mortensen, C. 1989. Mental images: Should cognitive science learn from neurophysiology? In (P. Slezak, ed) *_Computers, Brains and Minds_*. Kluwer.
- Pylyshyn, Z.W. 1973. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psych Bull* 80:1-24.
- Pylyshyn, Z.W. 1978. Imagery and artificial intelligence. In (W. Savage, ed) *_Perception and Cognition_*. University of Minnesota Press. Reprinted in (N. Block, ed) *_Readings in the Philosophy of Psychology_* (MIT Press, 1980).
- Pylyshyn, Z.W. 1981. The imagery debate: Analog media vs. tacit knowledge. *Psychological Review* 88:16-45. Reprinted in Block 1981.
- Reisberg, D. & Chambers, D. 1991. Neither pictures nor propositions: What can we learn from a mental image? *Canadian Journal of Psychology* 45:336-52.
- Rey, G. 1981. What are mental images? In (N. Block, ed) *_Readings in the Philosophy of Psychology_*, Vol. 2. Harvard University Press.
- Richardson, A. 1969. *_Mental Imagery_*. Routledge.
- Rollins, M. 1989. *_Mental Imagery: On the Limits of Cognitive Science_*. Yale University Press.
- Russow, L. 1985. Dennett, mental images and images in context. *Philosophy and Phenomenological Research* 45:581-94.
- Schwartz, R. 1980. Imagery: There is more to it than meets the eye. *Philosophy of Science Association* 1980.
- Shepard, R. & Cooper, L. 1982. *_Mental Images and their Transformations_*. MIT Press.
- Shier, D. 1997. How can pictures be propositions? *Ratio* 10:65-75.
- Sterelny, K. 1986. The imagery debate. *Philosophy of Science* 53:560-83. Reprinted in (W. Lycan, ed) *_Mind and Cognition_* (Blackwell, 1990).
- Thomas, N.J.T. 1997. Are theories of imagination theories of imagery? Manuscript.
- Tye, M. 1984. The debate about mental imagery. *Journal of Philosophy* 81:678-91.
- Tye, M. 1988. The picture theory of images. *Philosophical Review*.
- Tye, M. 1991. *_The Imagery Debate_*. MIT Press,
- Wright, E. 1983. Inspecting images. *Philosophy* 58:57-72.

5.1d Rationality

- Biro, J. & Ludwig, K. 1994. Are there more than minimal a priori limits on irrationality? *Australasian Journal of Philosophy* 72:89-102.
- Cherniak, C. 1986. *_Minimal Rationality_*. MIT Press.
- Cherniak, C. 1981. Minimal rationality. *Mind* 90:161-83.
- Cherniak, C. 1983. Rationality and the structure of memory. *Synthese* 57:163-86.
- Cohen, L.J. 1979. On the psychology of prediction: Whose is the fallacy? *Cognition* 7:385-407.
- Cohen, L.J. 1980. Whose is the fallacy? A rejoinder to Daniel Kahneman and Amos Tversky. *Cognition* 8:89-92.
- Cohen, L.J. 1981. Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*.
- Cohen, L.J. 1986. *_The Dialogue of Reason_*. Cambridge University Press.
- Cook, K.S. & Levi, M. 1990. *_The Limits of Rationality_*. University of Chicago Press.
- Davidson, D. 1985. Incoherence and irrationality. *Dialectica* 39:345-54.
- Davidson, D. 1995. Could there be a science of rationality? *International Journal of Philosophical Studies* 3:1-16.
- Feldman, R. 1988. Rationality, reliability, and natural selection. *Philosophy of Science* 55:218-27.
- Fetzer, J.H. 1990. Evolution, rationality and testability. *Synthese* 82:423-39.
- Gardner, S. 1996. *_Irrationality and the Philosophy of Psychoanalysis_*. Cambridge University Press.
- Harman, G. 1986. *_Change in View_*. MIT Press.
- Holt, L. 1999. Rationality is still hard work: Some further notes on the disruptive effects of deliberation. *Philosophical Psychology* 12:215-219.
- Kahneman, D., Slovic, P. & Tversky, A. (eds) 1982. *_Judgment under Uncertainty: Heuristics and Biases_*. Cambridge University Press.
- Kahneman, D. & Tversky, A. 1979. On the interpretation of intuitive probability: A reply to Jonathan Cohen. *Cognition* 7:409-11.
- Manktelow, K. & Over, D. 1987. Reasoning and rationality. *Mind and Language* 2:199-219.
- Mele, A.R. 1987. *_Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control_*. Oxford University Press.
- Nisbett, R. & Ross, L. 1980. *_Human Inference: Strategies and Shortcomings of Social Judgment_*. Prentice-Hall.
- Nozick, R. 1993. *_The Nature of Rationality_*. Princeton University Press.
- Reiner, R. 1995. Arguments against the possibility of perfect rationality. *Minds and Machines* 5:373-89.
- Rust, J. 1990. Delusions, irrationality, and cognitive science. *Philosophical*

Psychology.

Scholl, B.J. 1997. Reasoning, rationality, and architectural resolution. *Philosophical Psychology* 10:451-470.

Scott-Kakures, D. 1996. Self-deception and internal irrationality. *Philosophy and Phenomenological Research* 56:31-56.

Sober, E. 1981. The evolution of rationality. *Synthese* 46:95-120.

Stein, E. 1994. Rationality and reflective equilibrium. *Synthese* 99:137-72.

Stein, E. 1996. *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford University Press.

Stich, S.P. 1985. Could man be an irrational animal? *Synthese* 64:115-35.

Wason, P. 1966. Reasoning. In (Foss, ed) *New Horizons in Psychology*. Penguin.

5.1e Embodiment [see also 2.2]

Agre, P. 1995. Computation and embodied agency. *Informatica* 19:527-35.

Ballard, D. 1991. Animate vision. *Artificial Intelligence* 48:57-86.

Beer, R. 1995. A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* 72:173-215.

Bermudez, J.L., Marcel, A., & Eilan, N. (eds) 1995. *The Body and the Self*. MIT Press.

Buckley, J. & Hall, L. 1999. Self-knowledge and embodiment. *Southwest Philosophy Review* 15.

Chrisley, R.L. 1994. Taking embodiment seriously: Nonconceptual content and robotics. In (K.M. Ford, C. Glymour, & P. Hayes, eds) *Android Epistemology*. MIT Press.

Clark, A. 1987. Being there: Why implementation matters to cognitive science. *AI Review* 1:231-44.

On the importance of embodiment of systems in cognition.

Clark, A. 1995. Moving minds: Situating content in the service of real-time success. *Philosophical Perspectives* 9:89-104.

Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again*. MIT Press.

Clark, A. & Chalmers, D.J. 1998. The extended mind. *Analysis* 58:7-19.
Advocates a sort of "active externalism", based on the role of the environment in actively driving cognition. Beliefs can extend into an agent's immediate environment (e.g. a notebook) in this way.

Clark, A. 2001. Reasons, robots and the extended mind. *Mind and Language* 16:121-45.

Cussins, A. 1992. Content, embodiment, and objectivity: The theory of cognitive trails. *Mind* 101:651-88.

Gibson, J.J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Godfrey-Smith, P. 1996. Complexity and the Function of Mind in Nature. Cambridge University Press.

Haugeland, J. 1993. Mind embodied and embedded. In (Y. Houng & J. Ho, eds) Mind and Cognition:1993 International Symposium. Academia Sinica.
Argues that the mind is not just embedded but intimately intermingled with the world. With some systems-theoretic arguments arguing against a determinate interface. Mind is not an inner realm.

Hendriks-Jansen, H. 1996. Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought. MIT Press.

Hutchins, E. 1995. Cognition in the Wild. MIT Press.

Johnson, M.L. 1987. The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason. University of Chicago Press.

Johnson, M.L. 1995. Incarnate mind. *Minds and Machines* 5:533-45.

Kirsh, D. & Maglio, P. 1995. On distinguish epistemic from pragmatic action. *Cognitive Science* 18:513-49.

Loren, L.A. & Dietrich, E. 1997. Merleau-Ponty, embodied cognition, and the problem of intentionality. *Cybernetics and Systems* 28:345-58.

Losonsky, M. 1995. Emdedded systems vs. individualism. *Minds and Machines* 5:357-71.

McClamrock, R. 1995. Existential Cognition: Computational Minds in the World. University of Chicago Press.

O'Regan, K. 1992. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology* 46:461-88.

Rosenschein, S.J. & Kaelbling, L.P. 1995. A situated view of representation and control. *Artificial Intelligence* 73:149-73.

van Gelder, T. 1993. The distinction between mind and cognition. In (Y. Houng & J. Ho, eds) Mind and Cognition: 1993 International Symposium. Academia Sinica.
Argues against the contemporary "Cartesian" view of mind as an ontologically homogeneous inner representational realm that causes behavior, arguing for a holistic embodied view instead. Mind is therefore safe from elimination.

Varela, F., Thompson, E. & Rosch, E. 1991. The Embodied Mind: Cognitive Science and Human Experience. MIT Press.

Vera, A.H. & Simon, H.A. 1993. Situated action: A symbolic interpretation. *Cognitive Science* 17:7-48.

Wells, A. 1996. Situated action, symbol systems and universal computation. *Minds & Machines* 6:33-46.

Wilkerson, W.S. 1999. From bodily motions to bodily intentions: the perception of bodily activity. *Philosophical Psychology* 12:61-77.

Zhang, J. & Norman, D. 1994. Representations in distributed cognitive tasks. *Cognitive Science* 18:87-122.

5.1f Animal Cognition [see also 6.4c, 6.4d]

Allen, C. 1997. Animal cognition and animal minds. In (M. Carrier & P. Machamer, eds) Mindscapes: Philosophy, Science, and the Mind. Pittsburgh

University Press.

Allen, C. 1999. Animal concepts revisited: the use of self-monitoring as an empirical approach. *Erkenntnis* 51:537-544.

Allen, C. & Bekoff, M. 1992. On aims and methods of cognitive ethology. *Philosophy of Science Association* 1992, 2:110-24.

Allen, C. & Bekoff, M. 1995. Cognitive ethology and the intentionality of animal behavior. *Mind and Language* 10:313-328.

Allen, C. & Bekoff, M. 1997. *_Species of Mind: The Philosophy and Biology of Cognitive Ethology_*. MIT Press.

Allen, C. & Hauser, M. 1991. Concept attribution in nonhuman animals: Theoretical and methodological problems in ascribing complex mental processes. *Philosophy of Science* 58:221-40. Reprinted in Allen & Jamison 1996.

Bateson, P.P.G. & Klopfer, P.H. 1991. *_Perspectives in Ethology, Volume 9: Human Understanding and Animal Awareness_*. Plenum Press.

Beer, C.G. 1992. Conceptual issues in cognitive ethology. *Advances in the Study of Behavior* 21:69-109.

Bekoff, M. & Jamieson, D. (eds) 1996. *_Readings in Animal Cognition_*. MIT Press.

Bekoff, M. 1999. Social cognition: Exchanging and sharing information on the run. *Erkenntnis* 51:617-632.

Cheney, D.L. & Seyfarth, R.M. 1990. *_How Monkeys See the World: Inside the Mind of Another Species_*. University of Chicago Press.

Clark, S.R.L. 1987. The description and evaluation of animal emotion. In (C. Blakemore & S. Greenfield, eds) *_Mindwaves_*. Blackwell.

Cockburn, D. 1994. Human beings and giant squids (on ascribing human sensations and emotions to non-human creatures). *Philosophy* 69:135-50.

Crisp, R. 1996. Evolution and psychological unity. In (M. Bekoff & D. Jamieson, eds) *_Readings in Animal Cognition_*. MIT Press.

Davidson, D. 1982. Rational animals. *Dialectica* 36:317-28.

Dawkins, M.S. 1987. Minding and mattering. In (C. Blakemore & S. Greenfield, eds) *_Mindwaves_*. Blackwell.

Dawkins, M.S. 1990. From an animal's point of view: Motivation, fitness, and animal welfare. *Behavioral and Brain Sciences*.

Dennett, D.C. 1983. Intentional systems in cognitive ethology: The 'Panglossian paradigm' defended. *Behavioral and Brain Sciences* 6:343-90. Reprinted in *_The Intentional Stance_* (MIT Press, 1987).

Dennett, D.C. 1989. Cognitive ethology: Hunting for bargains or a wild goose chase? In (Montefiore, ed) *_Goals, No-Goals and Own Goals_*. Unwin Hyman.

Dennett, D.C. 1995. Do animals have beliefs? In (H. Roitblat & J. Meyer, eds) *_Comparative Approaches to Cognitive Science_*. MIT Press.

Dennett, D.C. 1996. *_Kinds of Minds_*. Basic Books.

Dreckmann, F. 1999. Animal beliefs and their contents. *Erkenntnis* 51:597-615.

- Dupre, J. 1996. The mental lives of nonhuman animals. In (M. Bekoff & D. Jamieson, eds) *_Readings in Animal Cognition_*. MIT Press.
- Fellows, R. 2000. Animal belief. *Philosophy* 75:587-599.
- Gaita, R. 1992. Animal thoughts. *Philosophical Investigations* 15:227-44.
- Gauker, C. 1990. How to learn language like a chimpanzee. *Philosophical Psychology* 4:139-46.
- Glock, H. 2000. Animals, thoughts and concepts. *Synthese* 123:35-104.
- Gould, J.L. & Gould, C.G. 1982. The insect mind: Physics or metaphysics? In (D. Griffin, ed) *_Animal Mind -- Human Mind_*. Springer-Verlag.
- Gould, J.L. & Gould, C.G. 1994. *_The Animal Mind_*. Scientific American Library.
- Griffin, D.R. (ed) 1982. *_Animal Mind -- Human Mind_*. Springer-Verlag.
- Griffin, D.R. 1984. *_Animal Thinking_*. Harvard University Press.
- Griffin, D.R. 1992. *_Animal Minds_*. University of Chicago Press.
- Harrison, P. 1991. Do animals feel pain? *Philosophy* 66:25-40.
- Heil, J. 1982. Speechless brutes. *Philosophy and Phenomenological Research* 42:400-406.
- Hendrichs, H. 1999. Different roots of human intentionality in mammalian mentality. *Erkenntnis* 51:649-668.
- Malcolm, N. 1973. Thoughtless brutes. *Proceedings and Addresses of the American Philosophical Association* 46:5-20.
- Nelson, J. 1983. Do animals propositionally know? Do they propositionally believe? *American Philosophical Quarterly* 20:149-60.
- Premack, D. & Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 4:515-629.
- Premack, D. 1986. *_Gavagai! or the Future History of the Animal Language Controversy_*. MIT Press.
- Proust, J. 1999. Mind, space and objectivity in non-human animals. *Erkenntnis* 51:545-562.
- Radner, D. 1993. Directed action and animal communication. *Ration* 6:135-54.
- Radner, D. 1999. Mind and function in animal communication. *Erkenntnis* 51:633-648.
- Ristau, C.A. (ed) 1991. *_Cognitive Ethology: The Minds of Other Animals_*. Lawrence Erlbaum.
- Roberts, R.C. 1996. Propositions and animal emotion. *Philosophy* 71:147-56.
- Routley, R. 1982. Alleged problems in attributing beliefs, and intentionality, to animals. *Inquiry* 24:385-417.
- Savage-Rumbaugh, E.S., Rumbaugh, D.M., & Boysen, S. 1980. Do apes use language? *American Scientist* 68:49-61.
- Savage-Rumbaugh, S. & Brakke, K.E. 1996. Animal language: Methodological and

interpretative issues. In (C. Allen & D. Jamison, eds) *_Readings in Animal Cognition_*. MIT Press.

Sebeok, T.A. & Umiker-Sebeok, J. 1980. *_Speaking of Apes: A Critical Anthology of Two-Way Communication with Man_*. Plenum Press.

Smit, H. 1995. Are animal displays bodily movements or manifestations of the mind? *Behavior and Philosophy* 23:13-19.

Sorabji, R. 1992. Animal minds. *Southern Journal of Philosophy* 31:1-18.

Stephan, A. 1999. Are animals capable of concepts? *Erkenntnis* 51:583-596.

Sterelny, K. 1990. Animals and individualism. In (P. Hanson, ed) *_Information, Language and Cognition_*. University of British Columbia Press.

Sterelny, K. 1995. Basic minds. *Philosophical Perspectives* 9:251-70.

Stich, S.P. 1978. Do animals have beliefs? *Australasian Journal of Philosophy* 57:15-28.

Wilder, H. 1996. Interpretative cognitive ethology. In (C. Allen & D. Jamison, eds) *_Readings in Animal Cognition_*. MIT Press.

Wilson, M.D. 1995. Animal ideas. *Proceedings and Addresses of the American Philosophical Association* 69:7-25.

5.2 Aspects of Mind

5.2a Pain and Pleasure

Aydede, M. 2000. An analysis of pleasure vis-a-vis pain. *Philosophy and Phenomenological Research* 61:537-570.

Beardman, S. 2000. The choice between current and retrospective evaluations of pain. 13:97-110.

Blum, A. 1991. A note on pleasure. *Journal of Value Inquiry* 25:367-70.

Chapman, C.R. 2000. Pain and folk theory. *Brain and Mind* 1:209-222.

Conee, E. 1984. A defense of pain. *Philosophical Studies* 46:239-48.

Cowan, J. 1968. *_Pleasure and Pain: A Study in Philosophical Psychology_*. Macmillan.

Dartnall, T. 2001. The pain problem. *Philosophical Psychology* 14:95-102.

Dennett, D.C. 1978. Why you can't make a computer that feels pain. *Synthese* 38. Reprinted in *_Brainstorms_* (MIT Press, 1978).

Douglas, G. 1998. Why pains are not mental objects. *Philosophical Studies* 91:127-148.

Edwards, R. 1975. Do pleasures and pains differ qualitatively? *Journal of Value Inquiry* 9:270-81.

Garfield, J.L. 2001. Pain deproblematized. *Philosophical Psychology* 14:103-7.

Gillett, G. 1991. The neurophilosophy of pain. *Philosophy* 66:191-206.

Goldstein, I. 1980. Why people prefer pleasure to pain. *Philosophy* 55.

- Goldstein, I. 1989. Pleasure and pain: unconditional intrinsic values. *Philosophy and Phenomenological Research*.
- Goldstein, I. 1999. Intersubjective properties by which we specify pain, pleasure, and other kinds of mental states. *Philosophy*.
- Graham, G. & Stephens, G. 1987. Minding your P's and Q's: Pain and sensible qualities. *Nous* 21:395-405.
- Grahek, N. 1991. Objective and subjective aspects of pain. *Philosophical Psychology* 4:249-66.
- Grahek, N. 1995. The sensory dimension of pain. *Philosophical Studies* 79:167-84.
- Gustafson, D. 1995. Belief in pain. *Consciousness and Cognition* 4:323-45.
- Gustafson, D. 2000. On the supposed utility of a folk theory of pain. *Brain and Mind* 1:223-228.
- Hall, R.J. 1989. Are pains necessarily unpleasant? *Philosophy and Phenomenological Research* 49:643-59.
- Hardcastle, V.G. 1997. When a pain is not. *Journal of Philosophy* 94:381-409.
- Hardcastle, V.G. 2000. *The Myth of Pain*. MIT Press.
- Kaufman, R. 1985. Is the concept of pain incoherent? *Southern Journal of Philosophy* 23:279-84.
- Langsam, H. 1995. Why pains are mental objects. *Journal of Philosophy* 6:303-13.
- Momeyer, R. 1975. Is pleasure a sensation? *Philosophy and Phenomenological Research* 36:113-21.
- Morris, K.J. 1996. Pain, injury, and first/third-person asymmetry. *Philosophy and Phenomenological Research* 56:125-56.
- Nelkin, N. 1986. Pains and pain sensations. *Journal of Philosophy* 83:129-48.
- Nelkin, N. 1994. Reconsidering pain. *Philosophical Psychology* 7:325-43.
- Newton, N. 1989. On viewing pain as a secondary quality. *Nous* 23:569-98.
- Pitcher, G. 1970. The awfulness of pain. *Journal of Philosophy* 48.
- Pitcher, G. 1970. Pain perception. *Philosophical Review* 74:368-93.
- Puccetti, R. 1975. Is pain necessary? *Philosophy* 50:259-69.
- Quinn, W. 1968. Pleasure -- disposition or episode? *Philosophy and Phenomenological Research* 28:578-86.
- Rachels, S. 2000. Is unpleasantness intrinsic to unpleasant experiences? *Philosophical Studies* 99:187-210.
- Rachlin, H. 1985. Pain and behavior. *Behavioral and Brain Sciences* 8:43-83.
- Resnik, D. 2000. Pain as a folk psychological concept: A clinical perspective. *Brain and Mind* 1:193-207.
- Sufka, K.J. & Lynch, M.P. 2000. Sensations and pain processes. *Philosophical Psychology* 13:299-311.

Tye, M. 1995. A representational theory of pains and their phenomenal character. *Philosophical Perspectives* 9:223-39.

Williams, B. 1959. Pleasure and belief. *Proceedings of the Aristotelian Society*.

5.2b Emotions

Addis, L. 1995. The ontology of emotion. *Southern Journal of Philosophy* 33:261-78.

Arregui, J.V. 1996. On the intentionality of moods: Phenomenology and linguistic analysis. *American Catholic Philosophical Quarterly* 70:397-411.

Bedford, E. 1957. Emotions. *Proceedings of the Aristotelian Society* 57:281-304.

Ben-Ze'ev, A. 1987. The nature of emotions. *Philosophical Studies* 52:393-409.

Ben-Ze'ev, A. 1990. Describing the emotions. *Philosophical Psychology* 3:305-17.

Ben-Ze'ev, A. 1992. Emotional and moral evaluations. *Metaphilosophy* 23:214-29.

Charland, L.C. 1995. Feeling and representing: Computational theory and the modularity of affect. *Synthese* 105:273-301.

D'Arms, J. & Jacobson, D. 2000. The moralistic fallacy: On the "appropriateness" of emotions. *Philosophy and Phenomenological Research* 61:65-90.

Davis, W. 1981. A theory of happiness. *American Philosophical Quarterly* 18:111-20.

DeLancey, C. 1997. Emotion and the computational theory of mind. In (S. O'Neillain, P. McKevitt, & E. MacAogain, eds) *Two Sciences of Mind*. John Benjamins.

DeLancey, C. 1998. Real emotions. *Philosophical Psychology* 11:467-487.

de Sousa, R. 1979. The rationality of emotions. *Dialogue*.

de Sousa, R. 1987. *The Rationality of Emotion*. MIT Press.

Deigh, J. 1994. Cognitivism in the theory of emotions. *Ethics* 104:824-54.

Goldie, P. 2000. *The Emotions: A Philosophical Exploration*. Oxford University Press.

Gordon, R.M. 1974. The aboutness of emotions. *American Philosophical Quarterly* 27:11-36.

Gordon, R.M. 1986. The passivity of emotions. *Philosophical Review* 95:339-60.

Gordon, R.M. 1987. *The Structure of Emotions: Investigations in Cognitive Philosophy*. Cambridge University Press.

Green, O. 1992. *The Emotions: A Philosophical Theory*. Kluwer.

Greenspan, P.S. 1988. *Emotions and Reasons: An Enquiry into Emotional Justification*. Routledge.

- Griffiths, P. 1989. Folk, functional and neurochemical aspects of mood. *Philosophical Psychology* 2:17-32.
- Haybron, D.M. 2001. Happiness and pleasure. *Philosophy and Phenomenological Research* 62:501-528.
- Helm, B.W. 1994. *The Significance of Emotions*. American Philosophical Quarterly 31:319-31.
- Irani, K.S. & Myers, G. 1983. *Emotion: Philosophical Studies*. Haven.
- Letwin, O. 1987. *Ethics, Emotion, and the Unity of the Self*. Croom Helm.
- Lormand, E. 1985. Toward a theory of moods. *Philosophical Studies* 47:385-407.
- Lyons, W. 1978. Emotions and behavior. *Philosophy and Phenomenological Research*.
- Marks, J. 1982. A theory of emotion. *Philosophical Studies* 42:227-42.
- McCullagh, C.B. 1990. The rationality of emotions and of emotional behavior. *Australasian Journal of Philosophy* 68:44-58.
- Morreal, J. 1983. Humor and emotion. *American Philosophical Quarterly* 20:297-304.
- Nash, R.A. 1989. Cognitive theories of emotion. *Nous* 23:481-504.
- Neu, J. 1971. *Emotion, Thought, and Therapy*. Cambridge University Press.
- Neu, J. 2000. *A Tear is an Intellectual Thing: The Meanings of Emotion*. Oxford University Press.
- Pugmire, D. 1994. Real emotion. *Philosophy and Phenomenological Research* 54:105-22.
- Rey, G. 1980. Functionalism and the emotions. In (A. Rorty, ed), *Explaining Emotions*. University of California Press.
- Roberts, R.C. 1995. Feeling one's emotions and knowing oneself. *Philosophical Studies* 77:319-38.
- Rorty, A.O. 1978. Explaining emotions. *Journal of Philosophy*.
- Rorty, A.O. (ed) 1980. *Explaining Emotions*. University of California Press.
- Rosenthal, D.M. 1983. Emotions and the self. In (K. Irani & G. Myers) *Emotion: Philosophical Studies*. Haven.
- Sizer, L. 2000. Towards a computational theory of mood. *British Journal for the Philosophy of Science* 51:743-770.
- Solomon, R.C. 1973. Emotion and choice. *Review of Metaphysics* 17:20-41.
- Taylor, G. 1975. Justifying the emotions. *Mind*.
- Thalberg, I. 1964. Emotion and thought. *American Philosophical Quarterly*.
- Wilkinson, S. 2000. Is 'normal grief' a mental disorder? *Philosophical Quarterly* 50:289-305.
- Wilson, J.R.S. 1972. *Emotion and Object*. Cambridge University Press.
- Wollheim, R. 1999. *On the Emotions*. Yale University Press.

5.2c Dreams [see also 6.2l]

Ayer, A. 1960. Professor Malcolm on dreams. *Journal of Philosophy*.

Chappell, V.C. 1963. The concept of dreaming. *Philosophical Quarterly* 13:193-213.

Chihara, C. 1965. What dreams are made of. *Theoria* 31:145-58.

Curley, E.M. 1975. Dreaming and conceptual revision. *Australasian Journal of Philosophy* 53:119-41.

Dennett, D.C. 1976. Are dreams experiences? *Philosophical Review* 73:151-71. Reprinted in *_Brainstorms_* (MIT Press, 1978).

Dunlop, C.E.M. (ed) 1977. *_Philosophical Essays on Dreaming_*. Cornell University Press.

Emmett, K. 1978. Oneiric experiences. *Philosophical Studies* 34:445-50.

Flanagan, O. 1995. Deconstructing dreams: The spandrels of sleep. *Journal of Philosophy* 92:5-27.

Flanagan, O. 1996. Self-expression in sleep: Neuroscience and dreams. In *_Self-Expressions_*. Oxford University Press.

Flanagan, O. 2000. *_Dreaming Souls: Sleep, Dreams, and the Evolution of the Conscious Mind_*. Oxford University Press.

Hunter, J. 1971. Some questions about dreaming. *Mind* 80:70-92.

Hunter, J. 1983. The difference between dreaming and being awake. *Mind* 92:80-93.

Landesman, C. 1964. Dreams: Two types of explanation. *Philosophical Studies* 15:17-23.

Malcolm, N. 1962. *_Dreaming_*. Routledge and Kegan Paul.

Mannison, D.S. 1975. Dreaming an impossible dream. *Canadian Journal of Philosophy* 4:663-75.

Matthews, G.B. 1981. On being immoral in a dream. *Philosophy* 56:47-64.

Putnam, H. 1962. Dreaming and 'depth grammar'. In (R. Butler, ed) *_Analytical Philosophy: First Series_*. Oxford University Press. Reprinted in *_Mind, Language, and Reality_* (Cambridge University Press, 1975).

Schroeder, S. 1997. The concept of dreaming: On three theses by Malcolm. *Philosophical Investigations* 20:15-38.

Seligman, M. & Yellen, A. 1987. What is a dream? *Behavior Research and Therapy* 25:1-24.

Shanon, B. 1983. Descartes' puzzle -- An organismic approach. *Cognition and Brain Theory* 6:185-95.

Siegler, F.A. 1967. Remembering dreams. *Philosophical Quarterly* 17:14-24.

5.2d Memory [see also 2.2f, 3.7]

Arcaya, J.M. 1989. Memory and temporality: A phenomenological alternative.

Philosophical Psychology 2:101-110.

Baier, A. 1976. Mixing memory and desire. American Philosophical Quarterly 13:213-20.

Ben-Zeev, A. 1986. Two approaches to memory. Philosophical Investigations 9:288-301.

Bergson, H. 1991. Matter and memory. MIT Press.

Campbell, J. 1997. The realism of memory. In (R. Heck, ed) *Language, Thought, and Logic: Essays in Honour of Michael Dummett*. Oxford University Press.

Campbell, J. 1997. The structure of time in autobiographical memory. European Journal of Philosophy 5:105-17.

Cherniak, C. 1983. Rationality and the structure of human memory. Synthese 57:163-86.

Furlong, E.J. 1956. The empiricist theory of memory. Mind 65:542-47.

Furlong, E.J. 1951. *A Study in Memory: A Philosophical Essay*. Nelson.

Gennaro, R.J. 1992. Consciousness, self-consciousness, and episodic memory. Philosophical Psychology 5:333-47.

Haight, D. & Haight, M. 1989. Time, memory, and self-remembering. Journal of Speculative Philosophy 3:1-11.

Holland, R.F. 1954. The empiricist theory of memory. Mind 63:464-86.

Judson, L. 1988. Russell on memory. Proceedings of the Aristotelian Society 88:65-82.

Kurtzman, H.S. 1983. Modern conceptions of memory. Philosophy and Phenomenological Research 44:1-20.

Malcolm, N. 1970. Memory and representation. Nous 4:59-71.

Martin, M.G.F. 1992. Perception, concepts, and memory. Philosophical Review 101:745-63.

Munsat, S. 1979. Memory and causality. In (D. Gustafson, ed) *Body, Mind, and Method*. Reidel.

Naylor, A. 1985. In defense of a nontraditional theory of memory. Monist 62:136-50.

Owens, D. 1996. A Lockean theory of memory experience. Philosophy and Phenomenological Research 56:319-32.

Perkins, R.K. 1973. Russell on memory. Mind 82:600-1.

Rakover, S. 1983. In defense of memory viewed as stored mental representation. Behaviorism 11:53-62.

Rosen, D.A. 1975. An argument for the logical notion of a memory trace. Philosophy of Science 42:1-10.

Rundle, B. 1986. Memory and causation. Philosophical Investigations 9:302-7.

Rychlak, J.F. 1996. Memory: A logical learning account. Journal of Mind and Behavior 17:229-50.

- Sanders, J.T. 1985. Experience, memory, and intelligence. *Monist* 68:507-21.
- Schectman, M. 1994. The truth about memory. *Philosophical Psychology* 7:3-18.
- Shope, R.K. 1973. Remembering, knowledge, and memory traces. *Philosophy and Phenomenological Research* 33:303-22.
- Stern, D.G. 1991. Models of memory: Wittgenstein and cognitive science. *Philosophical Psychology* 4:203-18.
- Sutton, J. 1998. *_Philosophy and Memory Traces: Descartes to Connectionism_*. Cambridge University Press.
- Urmson, J.O. 1971. Memory and imagination. *Mind* 80:607.
- Wilcox, S. & Katz, S. 1981. A direct realistic alternative to the traditional conception of memory. *Behaviorism* 9:227-40.
- Zemach, E. 1983. Memory: What it is, and what is cannot possibly be. *Philosophy and Phenomenological Research* 44:31-44.

5.2e Color [see also 1.3a, 1.7a, 1.7d]

- Armstrong, D.M. 1969. Colour realism and the argument from microscopes. In (R. Brown & C. Rollins, eds) *_Contemporary Philosophy in Australia_*. Humanities Press.
- Averill, E.W. 1985. Color and the anthropocentric problem. *Journal of Philosophy* 82:281-303.
- Averill, E.W. 1992. The relational nature of color. *Philosophical Review* 101:551-88.
- Bigelow, J. Collins, J. & Pargetter, R. 1990. Colouring in the world. *Mind* 99:279-88.
- Boghossian, P. & Velleman, J.D. 1989. Color as a secondary quality. *Mind* 98:81-103.
- Boghossian, P. & Velleman, J.D. 1991. Physicalist theories of color. *Philosophical Review* 100:67-106.
- Broackes, J. 1992. The autonomy of colour. In (D. Charles & K. Lennon, ed) *_Reduction, Explanation, and Realism_*. Oxford University Press.
- Broackes, J. 1997. *_The Nature of Colour_*. Routledge.
- Byrne, A. & Hilbert, D.R. 1997. *_Readings on Color, Volume 1: The Philosophy of Color_*. MIT Press.
- Byrne, A. & Hilbert, D.R. 1997. *_Readings on Color, Volume 2: The Science of Color_*. MIT Press.
- Byrne, A. & Hilbert, D.R. 1997. Colors and reflectances. In (A. Byrne & D.R. Hilbert, eds) *_Readings on Color, Volume 1: The Philosophy of Color_*. MIT Press.
- Campbell, J. 1993. A simple view of colour. In (J. Haldane & C. Wright, ed) *_Reality, Representation, and Projection_*. Oxford University Press.
- Campbell, K. 1969. Colours. In (R. Brown & C. Rollins, eds) *_Contemporary Philosophy in Australia_*. Humanities Press.

- Campbell, K. 1982. The implications of Land's theory of colour vision. In (L. Cohen, ed) *_Logic, Methodology, and Philosophy of Science_*, Vol. 6. North-Holland.
- Campbell, K. 1993. David Armstrong and realism about colour. In (J. Bacon, K. Campbell, & L. Reinhardt, eds) *_Ontology, Causality, and Mind_*. Cambridge University Press.
- Clark, A. 1996. True theories, false colors. *Philosophy of Science Supplement* 63:143-50.
- Dedrick, D. 1995. Objectivism and the evolutionary value of color vision. *Dialogue* 34:35-44.
- Dedrick, D. 1996. Can color be reduced to anything? *Philosophy of Science Supplement* 3:134-42.
- Foti, V.M. 1990. The dimension of color. *International Studies in Philosophy* 22:13-28.
- Gilbert, P. 1987. Westphal and Wittgenstein on white. *Mind* 76:399-403.
- Gilbert, P. 1989. Reflections on white: A rejoinder to Westphal. *Mind* 98:423-6.
- Gold, I. 1999. Dispositions and the central problem of color.
- Gold, I. 1999. On Lewis on naming the colours. *Australasian Journal of Philosophy* 77:365-370.
- Philosophical Studies* 93:21-44.
- Hall, R.J. 1996. The evolution of color vision without colors. *Philosophy of Science Supplement* 63:125-33.
- Hardin, C.L. 1983. Colors, normal observers and standard conditions. *Journal of Philosophy* 80:806-13.
- Hardin, C.L. 1984. A new look at color. *American Philosophical Quarterly* 21:125-33.
- Hardin, C.L. 1984. Are scientific objects colored? *Mind* 93:491-500.
- Hardin, C.L. 1985. The resemblances of colors. *Philosophical Studies* 48:35-47.
- Hardin, C.L. 1985. Frank talk about the colors of sense-data. *Australasian Journal of Philosophy* 63:485-93.
- Hardin, C.L. 1988. *_Color for Philosophers_*. Hackett.
- Hardin, C.L. 1988. Phenomenal colors and sorites. *Nous* 22:213-34.
- Hardin, C.L. 1989. Could white be green? *Mind* 390:285-8.
- Hardin, C.L. 1989. Idle colors and busy spectra. *Analysis* 49:47-8.
- Hardin, C.L. 1990. Color and illusion. In (W. Lycan, ed) *_Mind and Cognition_*. Blackwell.
- Hardin, C.L. 1993. van Brakel and the not-so-naked emperor. *British Journal for the Philosophy of Science* 44:137-50.
- Harvey, J. 1992. Challenging the obvious: The logic of color concepts.

Philosophia 21:277-94.

Harvey, J. 2000. Colour-dispositionalism and its recent critics. Philosophy and Phenomenological Research 61:137-156.

Hazen, A.P. 1999. On naming the colours. Australasian Journal of Philosophy 77:224-231.

Hilbert, D.R. 1987. _Color and Color Perception: A Study in Anthropocentric Realism_. CSLI Press.

Hilbert, D.R. 1992. What is color vision? Philosophical Studies 68:351-70.

Jackson, F. 1996. The primary quality view of color. Philosophical Perspectives 10:199-219.

Jackson, F. 1998. Colour, disjunctions, programming. Analysis 58:86-88.

Jackson, F. & Pargetter, R. 1987. An objectivist's guide to subjectivism about color. Revue Internationale de Philosophie 41:127-v41.

Jacovides, M. 2000. Cambridge changes of color. Pacific Philosophical Quarterly 81:142-164.

Johnston, M. 1992. How to speak of the colors. Philosophical Studies 68:221-263.

Kliewer, G. 1998. Neutral color concepts. Philosophical Studies 91:21-41.

Kraut, R. 1992. The objectivity of color and the color of objectivity. Philosophical Studies 3:265-87.

Langsam, H. 2000. Why colours do look like dispositions. Philosophical Quarterly 50:68-75.

Landesman, C. 1989. _Color and Consciousness: An Essay in Metaphysics_. Temple University Press.

Levin, J. 2000. Dispositional theories of color and the claims of common sense. Philosophical Studies 100:151-174.

Lewis, D. 1997. Naming the colours. Australasian Journal of Philosophy 75:325-42.

Maund, J.B. 1981. Colour: A case for conceptual fission. Australasian Journal of Philosophy 59:308-22.

Maund, J.B. 1991. The nature of color. History of Philosophy Quarterly 8:253-63.

Maund, J.B. 1995. _Colours: Their Nature and Representation_. Cambridge University Press.

McFarland, D. & Miller, A. 1998. Jackson on colour as a primary quality. Analysis 58:76-85.

McFarland, D. & Miller, A. 2000. Disjunctions, programming and the Australian view of colour. Analysis 60:209-212.

McGilvray, J.A. 1983. To color. Synthese 54:37-70.

McGilvray, J.A. 1994. Constant colors in the head. Synthese 100:197-239.

McGinn, C. 1996. Another look at color. Journal of Philoophy 93:537-53.

- McGinn, M. 1991. Westphal on the physical basis of color incompatibility. *Analysis* 4:218-22.
- McGinn, M. 1991. On two recent accounts of color. *Philosophical Quarterly* 41:316-24.
- Miller, A. 2001. The missing-explanation argument revisited. *Analysis* 61:76-86.
- Montgomery, R. 1996. The indeterminacy of color vision. *Synthese* 106:167-203.
- Nida-Rumelin, M. 1997. The character of color predicates: A phenomenalist view. In (M. Anduschus, A. Newen, & W. Kunne, eds) *Direct Reference, Indexicality, and Propositional Attitudes*. CSLI Press.
- Ross, P.W. 1999. The appearance and nature of color. *Southern Journal of Philosophy* 37:227-252.
- Ross, P. 2000. The relativity of color. *Synthese* 123:105-130.
- Smart, J.J.C. 1975. On some criticisms of a physicalist theory of colors. In (C. Cheng, ed) *Philosophical Aspects of the Mind-Body Problem*. University Press of Hawaii.
- Smart, J.J.C. 1995. 'Looks red' and dangerous talk. *Philosophy* 70_545-54.
- Smith, M.A. Color, transparency, mind-independence. In (J. Haldane & C. Wright, ed) *Reality, Representation, and Projection*. Oxford University Press.
- Smith, P. 1987. Subjectivity and colour vision. *Proceedings of the Aristotelian Society* 61:245-81.
- Spohn, W. 1997. The character of color predicates: A materialist view. In (M. Anduschus, A. Newen, & W. Kunne, eds) *Direct Reference, Indexicality, and Propositional Attitudes*. CSLI Press.
- Strawson, G. 1989. Red and 'red'. *Synthese* 78:193-232.
- Stroud, B. 2000. *The Quest for Reality: Subjectivism and the Metaphysics of Colour*. Oxford University Press.
- Stroud-Drinkwater, C. 1994. The naive theory of color. *Philosophy and Phenomenological Research* 54:345-54.
- Thompson, E., Palacios, A., & Varela, F.J. 1992. Ways of coloring. *Behavioral and Brain Sciences*.
- Thompson, E. 1995. Colour vision, evolution, and perceptual content. *Synthese* 104:1-32.
- Thompson, E. 1995. *Colour Vision*. Routledge.
- Tolliver, J.T. 1996. Interior colors. *Philosophical Topics* 22:411-41.
- van Brakel, J. 1993. The plasticity of categories: The case of color. *British Journal for the Philosophy of Science* 44:103-135.
- Watkins, M. 1999. Do animals see colors? An anthropocentrist's guide to animals, the color blind, and far away places. *Philosophical Studies* 94:189-209.
- Westphal, J. 1982. Brown: Remarks on color. *Inquiry* 25:417-33.

Westphal, J. 1986. White. Mind 95:310-28.

Westphal, J. 1989. Black. Mind 98:585-9.

Westphal, J. 1991. Colour: A Philosophical Introduction. Blackwell.

Whitmyer, VG. 1999. Ecological color. Philosophical Psychology 12:197-214.

Wittgenstein, L. 1977. Remarks on Colour. University of California Press.

5.3 Philosophy of Psychology, General

5.3a Psychological Laws [see also 3.5d]

Antony, L. 1995. Law and order in psychology. Philosophical Perspectives 9:429-46.

Braithwaite, M. 1949. Causal laws in psychology. Aristotelian Society Supplement 23:45-60.

Fodor, J.A. 1991. You can fool some of the people all of the time, everything else being equal: Hedged laws and psychological explanation. Mind 100:19-34.
Ceteris paribus means that every realizing state has completing conditions.
Even absolute exceptions are OK, as long as they're not across-the-board.

Fodor, J.A. 1989. Making mind matter more. Philosophical Topics 17:59-79.
Reprinted in A Theory of Content and Other Essays (MIT Press, 1990).
Non-strict psychological laws are compatible with the (nomologically sufficient) causal responsibility of mental properties. So there's no need for epiphobia. With comments on the relation between laws and mechanisms.

Horgan, T. & Tienson, J. 1990. Soft laws. Midwest Studies in Philosophy 15.
Argues that any laws in intentional psychology have ineliminable same-level exceptions; the Kuhnian crisis in cognitive science gives evidence for this.
But ceteris paribus laws provide perfectly good theoretical explanation.

Lycan, W.G. 1981. Psychological laws. Philosophical Topics 12:9-38.
A functionalist defense against anomalous monism. Psychofunctional laws and psychological laws, though not psychophysical laws, may exist. Rebutting arguments from rationality, indeterminism, intensionality, etc.

Mace, C.A. 1949. Causal laws in psychology. Aristotelian Society Supplement 23:61-68.

Marcello, G. 2000. Horgan and Tienson on ceteris paribus laws. Philosophy of Science 67:301-315.

Mott, P. 1992. Fodor and ceteris paribus laws. Mind 101:335-46.

Pietroski, P. & Rey, G. 1995. When other things aren't equal: Saving ceteris paribus laws from vacuity. British Journal for the Philosophy of Science 46:81-110.

Schiffer, S. 1991. Ceteris paribus laws. Mind 100:1-17.
There are no ceteris paribus laws, as there's no satisfactory way to cash the "unless" cause. But psychology doesn't need laws, anyway.

Silverberg, A. 1996. Psychological laws and nonmonotonic logic. Erkenntnis 44:199-224.

Warfield, T.A. 1993. Folk-psychological ceteris-paribus laws. Philosophical

Studies 71:99-112.

5.3b Psychology and Neuroscience [see also 6.1i]

Bechtel, W. 1983. A bridge between cognitive science and neuroscience: The functional architecture of mind. *Philosophical Studies* 44:319-30.

Arguing for the notion of functional architecture as a bridge whereby neural components can be components of cognitive processes.

Bub, J. 1994. Testing models of cognition through the analysis of brain-damaged patients. *British Journal for the Philosophy of Science* 45:837-55.

Butler, K. 1994. Neural constraints in cognitive science. *Minds and Machines* 4:129-62.

Cherniak, C. 1991. Meta-neuroanatomy: The myth of the unbounded mind/brain. In (E. Agazzi, ed) *_Philosophy and the Origin and Evolution of the Universe_*.

Cherniak, C. 1994. Philosophy and computational neuroanatomy. *Philosophical Studies* 73:89-107.

Argues that we can understand the brain under the hypothesis that it is optimized to "save wire", due to bounded resources: organization predicts placement. With remarks on the relation between cognitive and neural levels.

Churchland, P.M. 1986. Some reductive strategies in cognitive neurobiology. *Mind* 95:279-309. Reprinted in *_A Neurocomputational Perspective_* (MIT Press, 1989).

Some cute examples of neurophysiological reductions using state-spaces.

Churchland, P.M. 1995. *_The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain_*. MIT Press.

Churchland, P.S. 1980. A perspective on mind-brain research. *Journal of Philosophy* 77:185-207.

The brain can tell us a lot about the mind. With examples.

Churchland, P.S. 1982. Mind-brain reduction: New light from philosophy of science. *Neuroscience* 7:1041-7.

Churchland, P.S. 1986. *_Neurophilosophy: Toward A Unified Science of the Mind-Brain_*. MIT Press.

All about neuroscience, philosophy and prospects for their interaction.

Churchland, P.S. & Sejnowski, T. 1989. Neural representation and neural computation. In (L. Nadel, ed) *_Neural Connections, Mental Computations_*. MIT Press.

About how neuroscience and connectionism affect our conception of mind.

Churchland, P.S. 1987. Epistemology in the age of neuroscience. *Journal of Philosophy* 84:546-53.

On paradigm shifts, biology, evolution, connectionism, etc.

Clark, A. 1980. *_Psychological Models and Neural Mechanisms: An Examination of Reductionism in Psychology_*. Oxford University Press.

Glymour, C. 1994. On the methods of cognitive neuropsychology. *British Journal for the Philosophy of Science* 45:815-35.

Hardcastle, V.G. 1992. Reduction, explanatory extension, and the mind/brain sciences. *_Philosophy of Science_* 59:408-28.

The relationship between psychology and neuroscience is best characterized not by reduction but by explanatory extension, where each field is enriched by the other. With a number of examples from recent empirical work.

Hatfield, G. 1988. Neurophilosophy meets psychology: Reduction, autonomy, and empirical constraints. *Cognitive Neuropsychology* 5:723-46.

Hatfield, G. 2000. The brain's 'new' science: Psychology, neurophysiology, and constraint. *Philosophy of Science* 67:388-404.

Klagge, J.C. 1989. Wittgenstein and neuroscience. *Synthese* 78:319-43.
Wittgenstein wouldn't have liked the Churchlands, as neuro might be chaos, and too much neuro might undermine our self-conception nihilistically.

Kobes, B. 1991. On a model for psycho-neural coevolution. *Behavior and Philosophy* 19:1-17.

Madell, G. 1986. Neurophilosophy: A principled skeptic's response. *Inquiry*.

Manier, E. 1986. Problems in the development of cognitive neuroscience: Effective communication between scientific domains. *Philosophy of Science Association* 1986, 1:183-97.

McCauley, R. 1986. Intertheoretic relations and the future of psychology. *Philosophy of Science* 53:179-99.
Incommensurable theories don't necessarily require elimination, if their relationship is synchronic/interlevel, rather than diachronic/intralevel.

Mucciolo, L. 1974. The identity thesis and neuropsychology. *Nous* 8:327-42.
Argues contra Fodor and Block that neurological equipotentiality doesn't refute type materialism. Mental states may not be anatomically defined neural states, but they may be more abstract neural holograms.

Mundale, J. & Bechtel, W. 1996. Integrating neuroscience, psychology, and evolutionary biology through a teleological conception of function. *Minds and Machines* 6:481-505.

Ravenscroft, I. 1998. Neuroscience and the mind. *Mind and Language* 13:132-137.

Rockwell, W.T. 1994. On what the mind is identical with. *Philosophical Psychology* 7:307-23.
Argues that the mind is not identical with the brain -- at the very least, it's the central nervous system, and perhaps more. "Brain" does not denote a natural kind in neurophysiology.

Smith, A. 1986. Brain-mind philosophy. *Inquiry* 29:203-15.

Skarda, S. 1986. Explaining behavior: Bringing the brain back in. *Inquiry* 29:187-201.

Stoljar, D. & Gold, S. 1998. On biological and cognitive neuroscience. *Mind and Language* 13:110-31.

Stone, T. & Davies, M. 1993. Cognitive neuropsychology and the philosophy of mind. *British Journal for the Philosophy of Science* 44:589-622.

Stufflebeam, R.S. & Bechtel, W. 1997. PET: Exploring the myth and the method. *Philosophy of Science* 64:95-106.

van Orden, G.C. 1997. Functional neuroimages fail to discover pieces of mind in the parts of the brain. *Philosophy of Science Supplement* 64:85-94.

von Eckardt, B. 1984. Cognitive psychology and principled skepticism. *Journal*

of Philosophy 81:67-88.

Cognitive psychology can transmogrify itself, who needs neuroscience?

5.3c Explanation in Cognitive Science

Cummins, R. 1982. The internal manual model of psychological explanation. *Cognition and Brain Theory* 5:257-68.

Cummins, R. 1983. *The Nature of Psychological Explanation*. MIT Press.
Psychological explanation is typically via functional analysis, not causal subsumption. On interpretation, computation, and an analysis of cognition and intentionality. With remarks on Dretske, Searle, Titchener, Hull, Freud.

Fodor, J.A. 1968. *Psychological Explanation*. Random House.

Fodor, J.A. 1968. The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy* 65:627-40. Reprinted in *RePresentations* (MIT Press, 1980).

Franks, B. 1995. On explanation in cognitive science: Competence, idealization, and the failure of the classical cascade. *British Journal for the Philosophy of Science* 46:475-502.

Gilman, D. 1993. Optimization and simplicity: Marr's theory of vision and biological explanation. *Synthese* 107:293-323.
Contra Kitcher 1988, much of Marr's theory doesn't depend on optimization; in any case, optimization isn't so bad. With remarks on interdisciplinarity.

Heil, J. 1986. Formalism and psychological explanation. *Journal of Mind and Behavior* 7:1-10.

On the tension between formal explanation and representational explanation.

Kim, J. 1989. Mechanism, purpose, and explanatory exclusion. **Philosophical Perspectives** 3:77-108. Reprinted in *Supervenience and Mind* (Cambridge University Press, 1993).
Discusses the principle: there cannot be two independent explanations of the same phenomena. With application to purposive explanation of behavior, theory reduction, and eliminativism, and a discussion of explanatory realism.

Kim, J. 1990. Explanatory exclusion and the problem of mental causation. In (E. Villanueva, ed) *Information, Semantics, and Epistemology*. Blackwell.
On the problems posed by explanatory exclusion, and possible solutions. With focus on the problems as they arise for Dretske's and Davidson's theories.

Knight, D. 1997. A poetics of psychological explanation. *Metaphilosophy* 28:63-80.

Millikan, R.G. 1993. Explanation in biopsychology. In (J. Heil & A. Mele, eds) *Mental Causation*. Oxford University Press.

Montgomery, R. 1995. Explanation and evaluation in cognitive science. *Philosophy of Science* 62:261-82.

Montgomery, R. 1998. Grades of explanation in cognitive science. *Synthese* 114:463-495.

Morris, M. 1986. Causes of behavior. *Philosophical Quarterly* 36:123-44.

Moser, P. 1994. Naturalism and psychological explanation. *Philosophical Psychology* 7:63-84.

Owens, J. 1998. Psychological explanation and causal deviancy. *Synthese*

115:143-169.

Sober, E. 1978. Psychologism. Journal for the Theory of Social Behavior 8:165-91.

5.3d Philosophy of Cognitive Science, Misc

Bealer, G. 1987. The boundary between philosophy and cognitive science. Journal of Philosophy 86:553-55.

Philosophy is autonomous: empirical considerations can't affect it.

Bogdan, R. 2000. _Minding Minds: Evolving a Reflexive Mind by Interpreting Others_. MIT Press.

Fetzer, J.H. 1991. _Philosophy and Cognitive Science_. Paragon House.

Flanagan, O.J. 1984. _The Science of the Mind_. MIT Press.

Hardcastle, V. 1996. _How to Build a Theory in Cognitive Science_. SUNY Press.

Harnad, S. 1982. Neoconstructivism: A unifying constraint for cognitive science. In (T. Simon & R. Scholes, eds) _Language, Mind, and Brain_. Lawrence Erlbaum.

Haugeland, J. 1978. The nature and plausibility of cognitivism. Behavioral and Brain Sciences 1:215-26.

Hooker, C.A. 1975. The information-processing approach to the brain-mind and its philosophical ramifications. Philosophy and Phenomenological Research 36:1-15.

Keely, B. 2000. Neuroethology and the philosophy of cognitive science. Philosophy of Science 67:404-418.

Kukla, A. 1989. Non-empirical issues in psychology. American Psychologist 44:485-94.

On the role of non-empirical advances in psychology: e.g. in theory construction, coherence analysis, conceptual innovation, with the aid of logically necessary truths and the contingent/pragmatic a priori.

Lloyd, D. 1989. _Simple Minds_. MIT Press.

O'Neillain, S. 1995. _The Search for Mind: A New Foundation for Cognitive Science_. Ablex.

O'Neillain, S., McKevitt, P. & MacAogain, E. (eds) 1997. _Two Sciences of Mind_. John Benjamins.

Pickering, M. & Chater, N. 1995. Why cognitive science is not formalized folk psychology. Minds and Machines 5.

Preston, B. 1994. Behaviorism and mentalism: Is there a third alternative? Synthese 100:167-96.

van Gelder, T. 1998. The roles of philosophy in cognitive science. Philosophical Psychology 11:117-36.

von Eckardt, B. 1993. _What is Cognitive Science?_ MIT Press.

5.4 Philosophy of Mind, General

- Armstrong, D. 1999. *The Mind-Body Problem: An Opinionated Introduction*. Westview Press.
- Baker, L.R. 1989. Recent work in the philosophy of mind. *Philosophical Books* 30:1-9.
A general overview.
- Bealer, G. 1986. The logical status of mind. *Midwest Studies in Philosophy* 10.
- Bechtel, W. 1988. *Philosophy of Mind: An Overview for Cognitive Science*. Lawrence Erlbaum.
- Braddon-Mitchell, D. & Jackson, F. 1997. *Philosophy of Mind and Cognition*. Blackwell.
- Burge, T. 1992. Philosophy of language and mind: 1950-1990. *Philosophical Review* 100:3-52.
An overview of the last 40 years of the philosophy of language and the philosophy of mind, covering many issues and trends.
- Carruthers, P. 1986. *Introducing Persons: Theories and Arguments in the Philosophy of Mind*. SUNY Press.
- Churchland, P.M. 1984. *Matter and Consciousness*. MIT Press.
- Crane, T. 1995. *The Mechanical Mind*. Penguin.
- Dennett, D.C. 1978. Current issues in the philosophy of mind. *American Philosophical Quarterly* 15:249-261.
An overview of everything, circa 1978: logical behaviorism, functionalism, the identity theory, qualia, meaning, and so on, with bibliography.
- Graham, G. 1993. *Philosophy of Mind: An Introduction*. Blackwell.
- Guttenplan, S. 2000. *Mind's Landscape: An Introduction to the Philosophy of Mind*. Blackwell Publishers.
- Haldane, J.J. 1994. Analytical philosophy and the nature of mind: Time for another rebirth? In (R. Warner & T. Szubka, eds) *The Mind-Body Problem: A Guide to the Current Debate*. Blackwell.
- Haldane, J. 2000. The state and fate of contemporary philosophy of mind. *American Philosophical Quarterly* 37:301-21.
- Hannay, B. 1994. *Subjectivity and Reduction: An Introduction to the Mind-Body Problem*. Westview Press.
- Harman, G. 1989. Some philosophical issues in cognitive science. In (M. Posner, ed) *Foundations of Cognitive Science*. MIT Press.
- Kim, J. 1996. *Philosophy of Mind*. Westview Press.
- Lowe, E.J. 2000. *An Introduction to the Philosophy of Mind*. Cambridge University Press.
- McGinn, C. 1982. *The Character of Mind*. Oxford University Press.
- Phillips, H. 1995. *Vicissitudes of the I: An Introduction to the Philosophy of Mind*. Prentice-Hall.
- Quine, W.V. 1985. States of mind. *Journal of Philosophy* 82:5-8.
- Rey, G. 1997. *Contemporary Philosophy of Mind: A Contentiously Classical*

Approach_. Blackwell.

Rorty, R. 1982. Contemporary philosophy of mind. Synthese 53:323-48.
In praise of the "Ryle-Dennett" tradition, and the elimination of dualism
from the philosophy of mind.

Rorty, R. 1993. Consciousness, intentionality, and pragmatism. In (S.
Christensen & D. Turner, eds) _Folk Psychology and the Philosophy of Mind_.
Lawrence Erlbaum.
A pragmatist perspective on the recent history of the philosophy of mind,
focusing on consciousness, intentionality, and mental representation, and on
debates between Fodor, Dennett, Searle, Putnam, and Davidson.

Shaffer, J.A. 1964. _Philosophy of Mind_. Prentice-Hall.

Smith, P. & Jones, O. 1986. _The Philosophy of Mind: An Introduction_.
Cambridge University Press.

Sprague, E. 1999. _Persons and their Minds: A Philosophical Investigation_.
Westview Press.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

Part 6: Consciousness in Science [1599]

=====

- 6.1 Consciousness and Neuroscience [460]
 - 6.1a Neural Correlates of Consciousness [74]
 - 6.1b Neural Correlates of Visual Consciousness [28]
 - 6.1c Blindsight [50]
 - 6.1d Other Neuropsychological Disorders [71]
 - 6.1e Cerebral Hemispheres and Consciousness [45]
 - 6.1f Neural Timing and Consciousness (Libet) [19]
 - 6.1g Neural Synchrony and Binding [42]
 - 6.1h Consciousness and Anesthesia [44]
 - 6.1i Foundational Issues [31]
 - 6.1j Consciousness and Neuroscience, Misc [56]
- 6.2 Consciousness and Psychology [787]
 - 6.2a Cognitive Models of Consciousness [60]
 - 6.2b Unconscious Perception [73]
 - 6.2c Unconscious Processes [45]
 - 6.2d Visual Consciousness [37]
 - 6.2e Consciousness and Attention [35]
 - 6.2f Consciousness and Memory [84]
 - 6.2g Consciousness and Learning [33]
 - 6.2h Consciousness and Metacognition [22]
 - 6.2i Consciousness and Control [38]
 - 6.2j Consciousness and Imagery [24]
 - 6.2k Consciousness and Emotion [31]
 - 6.2l Consciousness, Sleep, and Dreaming [34]
 - 6.2m Consciousness in Psychoanalysis [57]
 - 6.2n Consciousness and Time [35]
 - 6.2o Self-Consciousness [46]
 - 6.2p Development of Consciousness [33]
 - 6.2q The Stream of Consciousness [29]
 - 6.2r Foundational Issues [19]
 - 6.2s Consciousness and Psychology, Misc [52]
- 6.3 Consciousness and Physics [77]
 - 6.3a Interpretation of Quantum Mechanics [27]
 - 6.3b Quantum Mechanisms of Consciousness [30]

- 6.3c Consciousness and Physics, Misc [20]
- 6.4 Consciousness and Science, Misc [275]
 - 6.4a Evolution of Consciousness [31]
 - 6.4b Consciousness and Language [23]
 - 6.4c Animal Consciousness [33]
 - 6.4d Animal Self-Consciousness [22]
 - 6.4e Altered States of Consciousness [27]
 - 6.4f Parapsychology and Consciousness [54]
 - 6.4g Phenomenology [34]
 - 6.4h Foundations [27]
 - 6.4i Miscellaneous [24]

6.1 Consciousness and Neuroscience

6.1a Neural Correlates of Consciousness

Alkire, M.T., Haier, R.J., & James, H.F. 1998. Toward the neurobiology of consciousness: Using brain imaging and anesthesia to investigate the anatomy of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Baars, B.J. & Newman, J. 1994. A neurobiological interpretation of the Global Workspace theory of consciousness. In (A. Revonsuo & M. Kamppinen, eds) *Consciousness in Philosophy and Cognitive Neuroscience*. Lawrence Erlbaum.

Baars, B.J. 1995. Surprisingly small subcortical structures are needed for the state of waking consciousness, while cortical projection areas seem to provide perceptual contents of consciousness. *Consciousness and Cognition* 4:159-62.

Baars, B.J., Newman, J. & Taylor, J.G. 1998. Neuronal mechanisms of consciousness: A relational global workspace approach. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Bogen, J.E. 1995. On the neurophysiology of consciousness, part I: An overview. *Consciousness and Cognition* 4:52-62.

Bogen, J.E. 1995. On the neurophysiology of consciousness, part II: Constraining the semantic problem. *Consciousness and Cognition* 4:137-58.

Bogen, J.E. 1997. Some neurophysiologic aspects of consciousness. *Seminars in Neurology* 17:95-103.

Bogen, J.E. 1998. Locating the subjectivity pump: The thalamic intralaminar nuclei. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Bremer, F. 1966. Neurophysiological correlates of mental unity. In (J. Eccles, ed) *Brain and Conscious Experience*. Springer.

Buser, P.A. & Rougeul-Buser, A. (eds) 1978. *Cerebral correlates of conscious experience*. Elsevier.

Coenen, A.M.L. 1998. Neuronal phenomena associated with vigilance and consciousness: From cellular mechanisms to electroencephalographic patterns. *Consciousness and Cognition* 7:42-53.

Crick, F. 1984. Functions of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences USA* 81:4586-93.

Crick, F. & Koch, C. 1990. Toward a neurobiological theory of consciousness.

Seminars in the Neurosciences 2:263-275.

Crick, F. & Koch, C. 1998. Consciousness and neuroscience. Cerebral Cortex.

Crick, F. & Koch, C. 2000. The unconscious homunculus. In (T. Metzinger, ed) Neural Correlates of Consciousness. MIT Press.

Damasio, A. 2000. A neurobiology for consciousness. In (T. Metzinger, ed) _Neural Correlates of Consciousness_. MIT Press.

Dimond, S. 1976. Brain circuits for consciousness. Brain, Behavior, and Evolution 13:376-95.

Duzel E., Yonelinas A.P., Mangun G.R., Heinze H.J., & Tulving E. 1997. Event-related brain potential correlates of two states of conscious awareness in memory. Proceedings of the National Academy of Sciences of the United States of America 94:5973-8.

Edelman, G. & Tononi, G. 2000. Reentry and the dynamic core: Neural correlates of conscious experience. In (T. Metzinger, ed) Neural Correlates of Consciousness. MIT Press.

Ellis, R. 2000. Efferent brain processes and the enactive approach to consciousness. Journal Of Consciousness Studies 7:40-50.

Flohr, H. 1990. Brain processes and phenomenal consciousness: A new and specific hypothesis. Theory and Psychology 1:245-62.

Flohr, H. 1992. Qualia and brain processes. In (A. Beckermann, H. Flohr, & J. Kim, eds) _Emergence or Reduction?: Prospects for Nonreductive Physicalism_. De Gruyter.

Flohr, H. 1995. Sensations and brain processes. Behavioral Brain Research 71:157-61.

Gallese, V. 2000. The acting subject: Toward the neural basis of social cognition. In (T. Metzinger, ed) _Neural Correlates of Consciousness_. MIT Press.

Gazzaniga, M. 1993. Brain mechanisms and conscious experience. In _Experimental and Theoretical Studies of Consciousness_ (Ciba Foundation Symposium 174). Wiley.

Gazzaniga, M. 1998. Brain and conscious experience. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) _Consciousness: At the Frontiers of Neuroscience_. Lippincott-Raven.

Gray, J.A. 1995. The contents of consciousness: A neuropsychological conjecture. Behavioral and Brain Sciences 18:659-76.

Greenfield, S. 1997. How might the brain generate consciousness? Communication and Cognition 30:285-300.

Greenfield, S. 1998. A rosetta stone for mind and brain? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness II_. MIT Press.

Grossenbacher, P.G. (ed) 1997. _Finding Consciousness in the Brain: A Neurocognitive Approach_. John Benjamins.

Goldman-Rakic, P.S. 1988. The prefrontal contribution to working memory and conscious experience. In (O. Creutzfeld & J. Eccles, eds) _The Brain and Conscious Experience_. Pontifical Academy.

Hobson, J.A. 1994. _The Chemistry of Conscious States_. Basic Books.

Hobson, J.A. 1997. Consciousness as a state-dependent phenomenon. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Jasper, H. 1998. Sensory information and conscious experience. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

John, E.R., Easton, P. & Isenhardt, R. 1997. Consciousness and cognition may be mediated by multiple independent coherent ensembles. *Consciousness and Cognition* 6:3-39.

Jones, B.E. 1998. The neural basis of consciousness across the sleep-waking cycle. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Jung, R. 1954. Correlation of bioelectrical and autonomic phenomena with alterations of consciousness and arousal in man. In (J. Delafresnaye, ed) *_Brain Mechanisms and Consciousness_*. Blackwell.

Kahn, D., Pace-Schott, E.F. & Hobson, J.A. 1997. Consciousness in waking and dreaming: The roles of neuronal oscillation and neuromodulation in determining similarities and differences. *Neuroscience* 78:13-38.

Kanwisher, N. 2001. Neural events and perceptual awareness. *Cognition* 79:89-113.

Kinsbourne, M. 1988. An integrated field theory of consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Kinsbourne, M. 1993. Integrated cortical field model of consciousness. In (Ciba Foundation) *_Experimental and Theoretical Studies of Consciousness_*. Wiley.

Kinsbourne, M. 1995. The intralaminar thalamic nuclei: Subjectivity pumps or attention-action co-ordinators? *Consciousness and Cognition* 4:167-71.

Kleitman, N. 1955. The role of the cerebral cortex in the development and maintenance of consciousness. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the Third Conference_*. Josiah Macy Foundation.

Koch, C. & Crick, F. 1994. Some further ideas regarding the neuronal basis of awareness. In (C. Koch & J. Davis, eds) *_Large-Scale Neuronal Theories of the Brain_*. MIT Press.

Koch, C. & Crick, F. 2000. Some thoughts on consciousness and neuroscience. In (M. Gazzaniga, ed) *_The New Cognitive Neurosciences: 2nd Edition_*. MIT Press.

Lehmann, D., Strik, W.K., Henggeler, B., Koenig, T. 1998. Brain electric microstates and momentary conscious mind states as building blocks of spontaneous thinking: I. Visual imagery and abstract thoughts. *International Journal of Psychophysiology* 29:1-11.

Libet, B. 1982. Brain stimulation in the study of neuronal functions for conscious sensory experiences. *Human Neurobiology* 1:235-42.

Libet, B. 1989. Conscious subjective experience vs. unconscious mental functions: A theory of the cerebral processes involved. In (R. Cotterill, ed) *_Models of Brain Function_*. Cambridge University Press.

Libet, B. 1996. Neural processes in the production of conscious experiences.

In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.

Libet, B. 1998. Do the models offer testable proposals of brain functions for conscious experience? In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Magoun, H.W. 1954. The ascending reticular system and wakefulness. In (J. Delafresnaye, ed) *_Brain Mechanism and Consciousness_*. Blackwell.

Markowitsch H.J. 1995. Cerebral bases of consciousness: A historical view. *Neuropsychologia* 33:1181-1192.

Metzinger, T. 2000. *_Neural Correlates of Consciousness: Empirical and Conceptual Questions_*. MIT Press.

Newman, J.B. 1995. Thalamic contributions to attention and consciousness. *Consciousness and Cognition* 4:172-93.

Newman, J.B. 1997. Putting the puzzle together: Toward a general theory of the neural correlates of consciousness. *Journal of Consciousness Studies* 4:47-66, 4:100-121.

Newman, J.B. & Baars, B.J. 1993. A neural attentional model for access to consciousness: A global workspace perspective. *Concepts in Neuroscience* 4:255-90.

O'Keefe, J. 1985. Is consciousness the gateway to the hippocampal cognitive map? A speculative essay on the neural basis of mind. In (D. Oakley, ed) *_Brain and Mind_*. Methuen.

Orpwood R.D. 1994. A possible neural mechanism underlying consciousness based on the pattern processing capabilities of pyramidal neurons in the cerebral cortex. *Journal of Theoretical Biology* 169:403-18.

Parvizi, J. & Damasio, A. 2001. Consciousness and the brainstem. *Cognition* 79:135-59.

Penfield, W. 1937. The cerebral cortex and consciousness. In *_The Harvey Lectures_*. Reprinted in (R. Wilkins, ed) *_Neurosurgical Classics_*. Johnson Reprint Corporation, 1965.

Penfield, W. 1955. The permanent record of the stream of consciousness. *Acta Psychologica* 11:47-69.

Raichle, M. 2000. The neural correlates of consciousness: An analysis of cognitive skill learning. In (M. Gazzaniga, ed) *_The New Cognitive Neurosciences: 2nd Edition_*. MIT Press.

Rapcsak, S. & Kaszniak, A. 2000. Searching for the neural correlates of consciousness: Clues from face recognition research. *Brain & Cognition* 42:37-40.

Rudell, A.P. & Hua, J. 1996. The recognition potential and conscious awareness. *Electroencephalography and Clinical Neurophysiology* 98:309-318.

Sewards, T. & Sewards, M. 2000. The awareness of thirst: Proposed neural correlates. *Consciousness & Cognition* 9:463-487.

Smythies, J. 1997. The functional neuroanatomy of awareness: With a focus on the role of various anatomical systems in the control of intermodal attention. *Consciousness and Cognition* 6:455-81.

Sokolov, E.N. 1992. The neurophysiological mechanisms of consciousness.

Journal of Russian and East European Psychology 30:6-12.

Strehler, B.L. 1991. Where is the self?: A neuroanatomical theory of consciousness. Synapse 7:44-91.

Stuss, D.T. 1991. Self, awareness, and the frontal lobes: A neuropsychological perspective. In (J. Strauss, ed) The Self: Interdisciplinary Approaches. Springer-Verlag.

Taylor, J.G. 1998. Cortical activity and the explanatory gap. Consciousness and Cognition 7:109-48.

Tononi, G. & Edelman, G. 1998. Consciousness and the integration of information in the brain. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) Consciousness: At the Frontiers of Neuroscience. Lippincott-Raven.

Umla, C. 2000. Conscious experience depends on multiple brain systems. European Psychologist 5:3-11.

Vanderwolf, C. 2000. Are neocortical gamma waves related to consciousness? Brain Research 855:217-224.

Verfaellie, M. & Keane, M.M. 1997. The neural basis of aware and unaware forms of memory. Seminars in Neurology 17:153-61.

Wolf, N.J. 1997. A possible role for cholinergic neurons of the basal forebrain and pontomesencephalon in consciousness. Consciousness and Cognition 6:574-596.

6.1b Neural Correlates of Visual Consciousness (see also 6.1c)

Anderson, R.A. 1997. Neural mechanisms in visual motion perception in primates. Neuron 18:865-872.

Cowey, A. 1996. Visual awareness: Still at sea with seeing? Current Biology 6:45-47.

Crick, F. & Koch, C. 1995. Cortical areas in visual awareness. Nature 377:294-5.

Crick, F. & Koch, C. 1995. Are we aware of neural activity in primary visual cortex? Nature 375:121-23.

Farah, M.J., O'Reilly, R.C. & Vecera, S.P. 1997. The neural correlates of perceptual awareness: Evidence from covert recognition in prosopagnosia. In (J. Cohen & J. Schooler, eds) Scientific Approaches to Consciousness. Lawrence Erlbaum.

Farah, M. 2000. The Cognitive Neuroscience of Vision. Blackwell Publishers.

Ffytche, D. 2000. Imaging conscious vision. In (T. Metzinger, ed) Neural Correlates of Consciousness. MIT Press.

Goodale, M.A. & Milner, A.D. 1992. Separate visual pathways for perception and action. Trends in Neuroscience 15:20-25.

Goodale, M. & Murphy, K. 2000. Space in the brain: Different neural substrates for allocentric and egocentric frames of reference. In (T. Metzinger, ed) Neural Correlates of Consciousness. MIT Press.

Hubel, D.H. 1998. Recordings from the striate cortex in awake behaving animals. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds)

Consciousness: At the Frontiers of Neuroscience. Lippincott-Raven.

Koch, C. 1995. Visual awareness and the thalamic intralaminar nuclei. *Consciousness and Cognition* 4:163-66.

Koch, C. 1996. Toward the neuronal substrate of visual consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Koch, C. 1998. The neuroanatomy of visual consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Koch, C. & Braun, J. 1996. Toward the neuronal correlate of visual awareness. *Current Opinion in Neurobiology* 6:158-64.

Leopold, D.A. & Logothetis, N.K. 1996. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* 379: 549-553.

Logothetis, N. & Schall, J. 1989. Neuronal correlates of subjective visual perception. *Science* 245:761-63.

Logothetis, N.K., Leopold, D.A. & Sheinberg, D.L. 1996. What is rivalling during binocular rivalry? *Nature* 30(6575):621-624.

Logothetis, N.K. & Leopold, D.A. 1998. Single-neuron activity and visual perception. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Logothetis, N.K. 1999. Binocular rivalry: A window onto consciousness. *Scientific American*.

Lumer, E. 2000. Binocular rivalry and human visual awareness. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Marzi, C., Girelli, M., Miniussi, C., Smania, N., & Maravita, A. 2000. Electrophysiological correlates of conscious vision: Evidence from unilateral extinction. *Journal of Cognitive Neuroscience* 12:869-877.

Milner, A.D. 1995. Cerebral correlates of visual awareness. *Neuropsychologia* 33:1117-30.

Milner, A.D. 1998. Streams and consciousness: Visual awareness and the brain. *Trends in Cognitive Sciences* 2:25-30.

Milner, A.D. & Goodale, M.A. 1995. *_The Visual Brain in Action_*. Oxford University Press.

Sewards, T. & Sewards, M. 2000. Visual awareness due to neuronal activities in subcortical structures: A proposal. *Consciousness & Cognition* 9:86-116

Sheinberg, D.L. & Logothetis, N.K. 1997. The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences USA* 94:3408-3413.

Stoerig P. & Cowey A. 1995. Visual perception and phenomenal consciousness. *Behavioural Brain Research* 71:147-156.

Vanni, S., Revonsuo, A., Saarinen, J. & Hari, R. 1996. Visual awareness of objects correlates with activity of right occipital cortex. *Neuroreport* 8:183-186.

Zeki, S. & Bartels, A. 1999. Toward a theory of visual consciousness.

Consciousness and Cognition 8:225-59.

6.1c Blindsight

Azzopardi, P., & Cowey, A. 1997. Is blindsight like normal, near-threshold vision? *Proceedings of the National Academy of Sciences USA* 94:14190.

Barbur, J.L., Watson, J.D.G., Frackowiak, R.D.G., & Zeki, S. 1993. Conscious visual perception without V1. *Brain* 116:1293-1302.

Braddick, O., Atkinson, J., Hood, B., Harkness, W. 1992. Possible blindsight in infants lacking one cerebral hemisphere. *Nature* 360:461-463.

Campion, J, Latto, R., & Smith, Y. 1983. Is blindsight an effect of scattered light, spared cortex, and near-threshold vision? *Behavioral and Brain Sciences* 6:423-86.

Carey, D.P., Goodale, M.A. & Sprowl, E.G. 1990. Blindsight in rodents: The use of a "high-level" distance cue in gerbils with lesions of primary visual cortex. *Behavioural Brain Research* 38:283-289.

Cowey, A. 1995. Blindsight in real sight. *Nature* 377:290-1.

Cowey, A. 1995. Blindsight in monkeys. *Nature* 373:247-9.

Cowey, A. & Stoerig, P. 1991. The neurobiology of blindsight. *Trends in Neurosciences* 14:140-5.

Cowey, A. & Stoerig, P. 1992. Reflections on blindsight. In (A. Milner & M. Rugg, eds) *_The Neuropsychology of Consciousness_*. Academic Press.

Cowey, A. & Stoerig, P. 1997. Visual detection in monkeys with blindsight. *Neuropsychologia* 35:929-39.

Danckert, J. & Goodale, M. 2000. Blindsight: A conscious route to unconscious vision. *Current Biology* 10:R64-R67.

Gazzaniga, M.S., Fendrich, R. & Wessinger, C.M. 1994. Blindsight reconsidered. *Current Directions in Psychological Science* 3:93-96.

Graves, R.E. & Jones, B.S. 1992. Conscious visual perceptual awareness vs. non-conscious visual spatial localisation examined with normal subjects using possible analogues of blindsight and neglect. *Cognitive Neuropsychology* 9:487-508.

Guzeldere, G., Flanagan, O., & Hardcastle, V. 2000. The nature and function of consciousness: Lessons from blindsight. In (M. Gazzaniga, ed) *_The New Cognitive Neurosciences: 2nd Edition_*. MIT Press.

Heywood, C.A., Cowey, A. & Newcombe, F. 1991. Chromatic discrimination in a cortically colour-blind observer. *European Journal of Neuroscience* 3:802-12.

Jackson, S. 2000. Perception, awareness and action: Insights from blindsight. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

Klein, S.A. 1998. Double-judgment psychophysics for research on consciousness: Application to blindsight. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Kolb, F.C. & Braun, J. 1995. Blindsight in normal observers. *Nature* 377:336-8.

Marcel, A.J. 1998. Blindsight and shape perception: Deficit of visual consciousness or of visual function? *Brain* 121:1565-88.

Marshall, J.C. & Halligan, P.W. 1988. Blindsight and insight in visuospatial neglect. *Nature* 336:766-67.

Morgan, M.J., Mason, A.J.S. & Solomon, J.A. 1997. Blindsight in normal subjects? *Nature* 385:401-2.

Natsoulas, T. 1982. Conscious perception and the paradox of "blind-sight". In (G. Underwood, ed) *_Aspects of Consciousness, Volume 3: Awareness and Self-Awareness_*. Academic Press.

Natsoulas, T. 1997. Blindsight and consciousness. *American Journal of Psychology* 110:1-33.

Paillard, J., Michel, F. & Stelmach, C.E. 1983. Localization without content: A tactile analogue of "blind sight". *Archives of Neurology* 40:548-51.

Place, U. 2000. Consciousness and the zombie within: A functional analysis of the blindsight evidence. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

Sahraie, A., Weiskrantz, L., Barbur, J.L., Simmons, A., & Brammer, M. 1997. Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Sciences USA* 94:9406-9411.

Stoerig, P. & Cowey, A. 1989. Wavelength sensitivity in blindsight. *Nature* 342:916-18.

Stoerig, P. & Cowey, A. 1989. Wavelength sensitivity in blindsight. Wavelength sensitivity in blindsight. *Brain* 115:425-44.

Stoerig, P. & Cowey, A. 1991. Increment threshold spectral sensitivity in blindsight: Evidence for colour opponency.

Stoerig, P. & Cowey, A. 1993. Blindsight and perceptual consciousness: Neuropsychological aspects of striate cortical function. In (B. Gulyas, D. Ottoson, & P. Roland, eds) *_Functional Organization of the Human Visual Cortex_*. Pergamon Press.

Stoerig, P. & Cowey, A. 1997. Blindsight in man and monkey. *Brain* 120:535-59.

Stoerig, P. 1998. Varieties of vision: From blind responses to conscious recognition. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Tapp, J.D. 1997. Blindsight in hindsight. *Consciousness and Cognition* 6:67-74.

Torjussen, T. 1978. Visual processing in cortically blind hemifields. *Neuropsychologia* 16:5-21.

Vision, G. 1998. Blindsight and philosophy. *Philosophical Psychology* 11:137-59.

Weiskrantz, L. 1986. *_Blindsight: A Case Study and Implications_*. Oxford University Press.

Weiskrantz, L. 1995. Blindsight: Not an island unto itself. *Current Directions in Psychological Science* 4:146-151.

Weiskrantz, L. 1995. Blindsight: Conscious vs. unconscious aspects. In (J. King & K. Pribram, eds) _Scale in Conscious Experience_. Lawrence Erlbaum.

Weiskrantz, L. 1996. Blindsight revisited. *Current Opinion in Neurobiology* 6:215-220.

Weiskrantz, L. 1997. _Consciousness Lost and Found_. Oxford University Press.

Weiskrantz, L. 1998. Consciousness and commentaries. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness II_. MIT Press.

Weiskrantz, L. & Cowey, A. 1970. Filling in the scotoma: A study of residual vision after striate cortex lesions in monkeys. (in (E. Stellar & J. Sprague, eds) _Progress in Physiological Psychology_. Academic Press.

Weiskrantz, L., Barbur, J.L. & Sahraie, A. 1995. Parameters affecting conscious versus unconscious visual discrimination without V1. *Proceedings of the National Academy of Sciences USA* 92:6122-26.

Weiskrantz, L. 2000. Blindsight: Implications for the conscious experience of emotion. In (R. Lane, L. Nadel, & G. Ahern, eds) _Cognitive Neuroscience of Emotion_. Oxford University Press.

Wessinger, C.M., Fendrich, R., Ptito, A., & Villemure, J. 1996. Residual vision with awareness in the field contralateral to a partial or complete functional hemispherectomy. *Neuropsychologia* 34:1129-1137.

Wessinger, C.M., Fendrich, R., Gazzaniga, M.S. 1997. Islands of residual vision in hemianopic patients. *Journal of Cognitive Neuroscience* 9:203-21.

Zeki S. 1996. The motion vision of the blind and the modularity of consciousness. *Transactions of the Medical Society of London* 112:11-18.

Zeki, S. & ffytche, D.H. 1998. The Riddoch syndrome: Insights into the neurobiology of conscious vision. *Brain* 121:25-45.

Zihl, J. 1980. "Blindsight": Improvement of visually guided eye movements by systematic practice in patients with cerebral blindness. *Neuropsychologia* 18:71-77.

Zihl, J. & von Cramon, D. 1980. Registration of light stimuli in the cortically blind hemifield and its effect on localization. *Behavior and Brain Research* 1:287-298.

Zihl, J. & Werth, R. 1984. Contributions to the study of "blindsight", parts I & II. *Neuropsychologia* 22:1-22.

6.1d Other Neuropsychological Disorders

Barr, W.B. 1998. Neurobehavioral disorders of awareness and their relevance to schizophrenia. In (X. Amador & A. David, eds) _Insight and Psychosis_. Oxford University Press

Bates, D. & Cartlidge, N. 1994. Disorders of consciousness. In (E. Critchley, ed) _The Neurological Boundaries of Reality_. Farrand.

Berti, A. & Rizzolatti, G. 1992. Visual processing without awareness: Evidence from unilateral neglect. *Journal of Cognitive Neuroscience* 4:345-51.

Bisiach, E., Luzzatti, C. & Perani, D. 1979. Unilateral neglect, representational schema, and consciousness. *Brain* 102:609-18.

Bisiach, E. & Rusconi, M.L. 1990. Breakdown of perceptual awareness in unilateral neglect. *Cortex* 26:643-49.

Bisiach, E. & Geminiani, G. 1991. Anosognosia related to hemiplegia and hemianopia. In (G. Prigatano & D. Schacter, eds) *_Awareness of Deficits after Brain Injury_*. Oxford University Press.

Bisiach, E. 1992. Understanding consciousness: Clues from unilateral neglect and related disorders. In (A. Milner & M. Rugg, eds) *_The Neuropsychology of Consciousness_*. Academic Press.

Bisiach, E. 1993. Mental representation in unilateral neglect and related disorders. *Quarterly Journal of Experimental Psychology* 46A:435-61.

Bisiach, E. & Berti, A. 1995. Consciousness in dyschiria. In (M. Gazzniga, ed) *_The Cognitive Neurosciences_*. MIT Press.

Brady, J.P. & Lind, D.L. 1961. Experimental analysis of hysterical blindness. *Archives of General Psychiatry* 4:331-39.

De Giorgio, C.M. & Lew, M.F. 1991. Consciousness, coma, and the vegetative state: Physical basis and definitional character. *Issues in Law and Medicine* 6:361-371.

de Haan, E.H.F., Young, A.W., & Newcombe, F. 1987. Face recognition without awareness. *Cognitive Neuropsychology* 4:385-415.

de Renzi, E. 1986. Current issues in prosopagnosia. In (H. Ellis, M. Jeeves, F. Newcombe, & A. Young, eds) *_Aspects of Face Processing_*. Martinus Nijhoff.

Driver, J. & Vuilleumier, P. 2001. Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition* 79:39-88.

Engelien, A., Huber, W., Silbersweig, D., Frith, C., & Frachowiak, R. 2000. The neural correlates of 'deaf-hearing' in man. *Brain* 123:532-545.

Farah, M.J. 1990. *_Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision_*. MIT Press.

Farah, M.J. 1994. Perception and awareness after brain damage. *Current Opinion in Neurobiology* 4:252-55.

Farah, M.J. 1994. Visual perception and visual awareness after brain damage: A tutorial overview. In (C. Umiltà and M. Moscovitch, eds) *_Consciousness and Unconscious Information Processing: Attention and Performance 15_*. MIT Press.

Farah, M.J., O'Reilly, R.C. & Vecera, S.P. 1997. The neural correlates of perceptual awareness: Evidence from covert recognition in prosopagnosia. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Farah, M.J. & Feinberg, T.E. 1997. Consciousness of perception after brain damage. *Seminars in Neurology* 17:145-52.

Farah, M. & Feinberg, T. 2000. Disorders of perception and awareness. In (M. Farah & T. Feinberg, eds) *_Patient-based Approaches to Cognitive Neuroscience_*. MIT Press.

Farah, M. 2001. Consciousness. In (B. Rapp, ed) *_The Handbook of Cognitive Neuropsychology: What Deficits Reveal about the Human Mind_*. Psychology Press/Taylor & Francis.

Feinberg, T.E. 1997. Some interesting perturbations of the self in neurology. *Seminars in Neurology* 17:129-35.

- Fredericks, J.A.M. 1969. Consciousness. In (P. Vinken & G. Bruyn, eds) *Handbook of Clinical Neurology*. North Holland.
- Frith, C., Blakemore, S.J., & Wolpert, D. 2000. Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews* 31:357-363.
- Galin, D. 1992. Theoretical reflections on awareness, monitoring, and self in relation on anosognosia. *Consciousness and Cognition* 1:152-62.
- Giacino J.T. 1997. Disorders of consciousness: differential diagnosis and neuropathologic features. *Seminars in Neurology* 17:105-11.
- Gibson, K.R. 1992. Toward an empirical basis for understanding consciousness and self-awareness. *Consciousness and Cognition* 1:163-68.
- Grosz, H.J. & Zimmerman, J.A. 1965. Experimental analysis of hysterical blindness: A follow-up report and new experimental data. *Archives of General Psychiatry* 13:255-60.
- Hellman, K.M. 1991. Anosognosia: Possible neuropsychological mechanisms. In (G. Prigatono & D. Schacter, eds) *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Oxford University Press.
- Humphreys, G.W., Troscianko, T., Riddoch, M.J., & Boucart, M. 1992. Covert processing in different visual recognition systems. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.
- Jehkonen, M., Ahonen, J., Dastidar, P., & Vilkki, J. 2000. Unawareness of deficits after right hemisphere stroke: Double-dissociations of anosognosias. *Acta Neurologica Scandinavica* 102:378-384.
- Jouvet, M. 1969. Coma and other disorders of consciousness. In (P. Vinken & G. Bruyn, eds) *Handbook of Clinical Neurology*. North Holland.
- Katz, J. 2000. Individual differences in the consciousness of phantom limbs. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.
- Kihlstrom, J.F. & Tobias, B.A. 1991. Anosognosia, consciousness, and the self. In (G. Prigatono & D. Schacter, eds) *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Oxford University Press.
- Knight, R.T. & Grabowecky, M. 1995. Escape from linear time: Prefrontal cortex and conscious experience. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.
- Koehler, S. & Moscovitch, M. 1997. Unconscious visual processing in neuropsychological syndromes: A survey of the literature and evaluation of models of consciousness. In (M. Rugg, ed) *Cognitive Neuroscience*. MIT Press.
- Ladavas, E., Berti, A., & Farne, A. 2000. Dissociation between conscious and non-conscious processing in neglect. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.
- Lane, R.D., Ahern, G.L., Schwartz, G.E. & Kaszniak, A.W. 1997. Is alexithymia the emotional equivalent of blindsight? *Biological Psychiatry* 42:834-44.
- Light, G., & Braff, D. 2000. Do self-reports of perceptual anomalies reflect gating deficits in schizophrenia patients? *Biological Psychiatry* 47:463-467.

- McGlynn, S.M. & Schacter, D.L. 1989. Unawareness of deficits in neuropsychological syndromes. *Journal of Clinical and Experimental Neuropsychology* 11:143-205.
- Milner, A.D. & Rugg, M. (eds) 1991. *The Neuropsychology of Consciousness*. Academic Press.
- Milner, A.D. 1991. Disorders of perceptual awareness: Commentary. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.
- Newcombe, F. 1985. Neuropsychology of consciousness: A review of human clinical evidence. In (D. Oakley, ed) *Brain and Mind*. Methuen.
- Porter R.J. 1991. Disorders of consciousness and associated complex behaviors. *Seminars in Neurology* 11:110-17.
- Posner J.B. 1978. Coma and other states of consciousness: the differential diagnosis of brain death. *Annals of the New York Academy of Sciences* 315:215-27.
- Prigatono, G.P. & Schacter, D.L. (eds) 1991. *Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues*. Oxford University Press.
- Ramachandran, V.S. 1995. Anosognosia in parietal lobe syndrome. *Consciousness and Cognition* 4:22-51.
- Rioch, D.M. 1954. Psychopathological and neuropathological aspects of consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.
- Schacter, D.L., McAndrews, M.P., and Moscovitch, M. 1986. Access to consciousness: Dissociations between implicit and explicit knowledge in neuropsychological syndromes. In (L. Weiskrantz, ed) *Thought Without Language*. Oxford University Press.
- Schacter, D.L. 1990. Toward a cognitive neuropsychology of awareness: Implicit knowledge and anosognosia. *Journal of Clinical and Experimental Neuropsychology* 12:155-78.
- Schiff, N. & Plum, F. 2000. The role of arousal and "gating" systems in the neurology of impaired consciousness. *Journal Of Clinical Neurophysiology* 17:438-452.
- Teasdale G., Knill-Jones R., & van der Sande J. 1978. Observer variability in assessing impaired consciousness and coma. *Journal of Neurology, Neurosurgery and Psychiatry* 41:603-10.
- Tononi, G. & Edelman, G. 2000. Schizophrenia and the mechanisms of conscious integration. *Brain Research Reviews* 31:391-400.
- Tranel, D. & Damasio, A.R. 1988. Nonconscious face recognition in patients with prosopagnosia. *Behavioral Brain Research* 30:235-49.
- Vaina, L.M. 1995. Akinetopsia, achromatopsia and blindsight: Recent studies on perception without awareness. *Synthese* 105:253-271.
- van de Kelt E., Segnarbieux F., Candon E., Couchet P., Frerebeau P., Daures J.P. 1994. Clinical recovery of consciousness after traumatic coma. *Critical Care Medicine* 22:1108-13.
- von Cramon, D. 1978. Consciousness and disturbances of consciousness. *Journal of Neurology* 219:1-13.

Vecera, S.P. & Gilds, K.S. 1997. What is it like to be a patient with apperceptive agnosia? *Consciousness and Cognition* 6:237-66.

Watson, R.T., Valenstein, E., Day, A., & Heilman, K.M. 1994. Posterior neocortical systems subserving awareness and neglect: Neglect associated with superior temporal sulcus but not area 7 lesions. *Archives of Neurology* 51:1014-1021

Weiskrantz, L. 1987. Neuropsychology and the nature of consciousness. In (C. Blakemore & S. Greenfield, eds) *Mindwaves*. Blackwell.

Weiskrantz, L. 1988. Some contributions of neuropsychology of vision and memory to the problem of consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Weiskrantz, L. 1990. Outlooks for blindsight: Explicit methodologies for implicit processes. *Proceedings of the Royal Society of London* B239:247-78.

Weiskrantz, L. 1994. Neuropsychology and the nature of consciousness. In (H. Gutfreund & G. Toulouse, eds) *Biology and Computation: A Physicist's Choice*. World Scientific.

Yamadori, A. 1997. Body awareness and its disorders. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *Cognition, Computation, and Consciousness*. Oxford University Press.

Young, A.W. 1994. Covert recognition. In (M. Farah & G. Ratcliff, eds) *The Neuropsychology of High-Level Vision*. Lawrence Erlbaum.

Young, A.W. 1994. Conscious and unconscious recognition of familiar faces. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance* 15. MIT Press.

Young, A.W. 1995. Neuropsychology of awareness. In (A. Revonsuo & M. Kampinnen, eds) *Consciousness in Philosophy and Cognitive Neuroscience*. Lawrence Erlbaum.

Young, A.W. 1995. Face recognition and awareness after brain injury. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Young, A.W. 1996. Dissociable aspects of consciousness. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Zappulla R.A. 1997. Epilepsy and consciousness. *Seminars in Neurology* 17:113-19.

6.1e Cerebral Hemispheres and Consciousness

Albert M.L., Silverberg R., Reches A., & Berman M. 1976. Cerebral dominance for consciousness. *Archives of Neurology* 33:453-4.

Austin, G., Hayward, W., & Rouhe, S. 1974. A note on the problem of conscious man and cerebral disconnection by hemispherectomy. In (M. Kinsbourne & W. Smith, eds) *Hemispheric Disconnection and Cerebral Function*. Charles C. Thomas.

Battro, A. 2001. *Half a Brain is Enough: The Story of Nico*. Cambridge University Press.

Baynes, K. & Gazzaniga, M. 2000. Consciousness, introspection, and the split-brain: The two minds/one body problem. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.

- Beaumont, J. 1981. Split brain studies and the duality of consciousness. In (G. Underwood & R. Stevens, eds) *_Aspects of Consciousness, Volume 2_*. Academic Press.
- Bogen, J.E. 1968. The other side of the brain: An appositional mind. *Bulletin of the Los Angeles Neurological Society* 34:135-62.
- Bogen, J.E. 1977. Further discussion of split brains and hemispheric capabilities. *British Journal for the Philosophy of Science* 28:281-6.
- Dewitt, L. 1975. Consciousness, mind, self: The implications of the split-brain studies. *British Journal for the Philosophy of Science* 27:41-47.
- Dimond, S.J. 1978. Depletion of awareness and double-simultaneous stimulation in split-brain man. *Cortex* 14:604-607.
- Gazzaniga, M. 1977. On dividing the self: Speculations from brain research. *Excerpta Medica: Neurology* 434:233-44.
- Gazzaniga, M.S., LeDoux, J.E., Wilson, D.H. 1977. Language, praxis, and the right hemisphere: Clues to some mechanisms of consciousness. *Neurology* 27:1144-1147.
- Gazzaniga, M.S. 1995. Consciousness and the cerebral hemispheres. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.
- Gazzaniga, M. & Miller, M. 2000. Testing Tulving: The split brain approach. In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.
- Harrington, A. 1985. Nineteenth-century ideas on hemisphere differences and "duality of mind." *Behavioral and Brain Sciences* 8:617-660.
- Joseph, R. 1988. The right cerebral hemisphere: Emotion, music, visual-spatial skills, body-image, dreams, and awareness. *Journal of Clinical Psychology* 44:630-673.
- Kavcic, V., Fei, R., Hu, S., & Doty, R. 2000. Hemispheric interaction, metacontrol, and mnemonic processing in split-brain macaques. *Behavioural Brain Research* 111:71-82.
- Kurian, G., Santhakumari, K. 1990. Consciousness and the left cerebral hemisphere. *Journal of Indian Psychology* 8:33-36.
- LeDoux, J.E., Wilson, D.H. & Gazzaniga, M.S. 1977. A divided mind: Observations of the conscious properties of the separated hemispheres. *Annals of Neurology* 2:417-21.
- LeDoux, J.E., Wilson, D.H., & Gazzaniga, M.S. 1979. Beyond commissurotomy: Clues to consciousness. In (M. Gazzaniga, ed) *_Handbook of Behavioral Neurobiology_*, volume 2. Plenum Press.
- LeDoux, J.E. 1986. Brain, mind, and language. In (D. Oakley, ed) *_Brain and Mind_*. Methuen.
- Landis, T., Graves, R., & Goodglass, H. 1981. Dissociated awareness of manual performance on two different visual associative tasks: A "split-brain" phenomenon in normal subjects? *Cortex* 17:435-440.
- Lishman, W.A. 1971. Emotion, consciousness, and will after brain bisection in man. *Cortex* 7:181-92.
- Mackay, D.M. 1987. Divided brains -- divided minds? In (C. Blakemore and S.

Greenfield, eds) *Mindwaves*. Blackwell.

Mark, V. 1996. Conflicting communication in a split-brain patient: Support for dual consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Marks, C. 1980. *Commissurotomy, Consciousness, and Unity of Mind*. MIT Press.

Miller, L. 1986. Some comments on cerebral hemispheric models of consciousness. *Psychoanalytic Review* 73:129-44.

Natsoulas, T. 1987. Consciousness and commissurotomy: 1. Spheres and Streams of consciousness. *Journal of Mind and Behavior* 8:435-468.

Natsoulas, T. 1988. Consciousness and commissurotomy: 2. Some Pertinencies for Intact Functioning. *Journal of Mind and Behavior* 9:515-548.

Natsoulas, T. 1991. Consciousness and commissurotomy: 3. Toward the improvement of alternative conceptions. *Journal of Mind and Behavior* 12:1-32.

Natsoulas, T. 1992. Consciousness and commissurotomy: 4. Three hypothesized dimensions of deconnected left-hemispheric consciousness. *Journal of Mind and Behavior* 13:37-67.

Natsoulas, T. 1991. Consciousness and commissurotomy: 5. Concerning a hypothesis of normal dual consciousness. *Journal of Mind and Behavior* 14:179-202.

Natsoulas, T. 1991. Consciousness and commissurotomy: 6. Evidence for normal dual consciousness. *Journal of Mind and Behavior* 16:181-205.

Preilowski B. 1979. Self-recognition as a test of consciousness in left and right hemisphere of "split-brain" patients. *Activitas Nervosa Superior* 19 (supp):343-44.

Puccetti, R. 1977. Bilateral organization of consciousness in man. *Annals of the New York Academy of Sciences* 299:448-58.

Puccetti, R. 1981. The case for mental duality: Evidence from split-brain data and other considerations. *Behavioral and Brain Sciences* 4:93-123.

Quen, J.M. (ed) 1986. *Split Minds/Split Brains: Historical and Current Perspectives*. New York University Press.

Sergent, J. 1987. A new look at the human split brain. *Brain* 110:1375-92.

Sperry, R.W. 1968. Hemisphere deconnection and unity in conscious awareness. *American Psychologist* 23:723-733.

Sperry, R.W. 1977. Forebrain commissurotomy and conscious awareness. *Journal of Medicine and Philosophy* 2:101-26.

Sperry, R.W., Zaidel, E., Zaidel, D. 1979. Self recognition and social awareness in the deconnected minor hemisphere. *Neuropsychologia* 17:153-166.

Sperry, R.W. 1984. Consciousness, personal identity and the divided brain. *Neuropsychologia* 22:611-73.

Trevarthen, C. 1974. Analysis of central activities that generate and regulate consciousness in commissurotomy patients. In (S. Dimond & J. Beaumont, eds) *Hemisphere Function in the Human Brain*. Elek.

Wessinger, C.M., Fendrich, R., Ptito, A., & Villemure, J.G. 1996. Residual

vision with awareness in the field contralateral to a partial or complete functional hemispherectomy. *Neuropsychologia* 34:1129-1137.

Wilkes, K.V. 1978. Consciousness and commissurotomy. *Philosophy* 53:185-99.

Zangwill, O.L. 1974. Consciousness and the cerebral hemispheres. In (S. Dimond & J. Beaumont, eds) *Heremisphere Function in the Human Brain*. Wiley.

6.1f Neural Timing and Consciousness (Libet, etc)

Churchland, P.S. 1981. On the alleged backward referral of experience and its relevance to the mind-body problem. *Philosophy of Science* 48:165-81.

Churchland, P.S. 1981. The timing of sensations: Reply to Libet. *Philosophy of Science* 48:492-7.

Dennett, D.C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15:183-201.

Elitzur, A. 1996. Time and consciousness: The uneasy bearing of relativity on the mind-body problem. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Glynn, I.M. 1990. Consciousness and time. *Nature* 348:477-79.

Green, C. & Gillett, G. 1995. Are mental events preceded by their physical causes? *Philosophical Psychology* 8:333-340.

Honderich, T. 1984. The time of a conscious sensory experience and mind-brain theories. *Journal of Theoretical Biology* 110:115-129.

Kiefer, M. & Spitzer, M. 2000. Time course of conscious and unconscious semantic brain activation. *Neuroreport* 11:2401-2407.

Libet, B. 1978. Neuronal vs. subjective timing for a conscious sensory experience. In (P. Buser & A. Rougeul-Buser, eds) *Cerebral Correlates of Conscious Experience*. Elsevier.

Libet, B. Wright, E.W, Feinstein, B. & Pearl, D.K. 1979. Subjective referral of the timing for a cognitive sensory experience. *Brain* 102:193-224.

Libet, B. 1981. The experimental evidence for subjective referral of a sensory experience backwards in time: Reply to P.S. Churchland. *Philosophy of Science*, 48, 182-97.

Libet, B. 1981. Timing of cerebral processes relative to concomitant conscious experiences in man. In (G. Adam, I. Meszaros & E.I. Banyai, eds), *Advances in Physiological Science*. Pergamon.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8:529-66.

Libet, B. 1985. Subjective antedating of a sensory experience and mind-brain theories: Reply to Honderich. *Journal of Theoretical Biology* 114:563-70.

Libet, B., Wright, E.W., Feinstein, B. & Pearl, D.K. 1992. Retroactive enhancement of a skin sensation by a delayed cortical stimulus in man: Evidence for delay of a conscious sensory experience. *Consciousness and Cognition* 1:367-75.

Libet, B. 1993. The neural time factor in conscious and unconscious events. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

Libet, B. 1993. *Neurophysiology of Consciousness: Selected Papers and New Essays*. Birkhauser.

Mele, A.R. 1997. Strength of motivation and being in control - Learning from Libet. *American Philosophical Quarterly* 34:319-32.

Rossi, E.L. 1988. Paradoxes of time, consciousness, and free will: Integrating Bohm, Jung, and Libet on ethics. *Psychological Perspectives* 19:50-55.

6.1g Neural Synchrony and Binding

Cotterill, R.M. & Nielsen, C. 1991. A model for cortical 40-hertz oscillations invokes inter-area interactions. *Neuroreport* 2:289-92.

Crick, F. & Koch, C. 1990. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2:263-275.

Damasio, A.R. 1989. The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation* 1:123-32.

Damasio, A.R. 1989. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recognition and recall. *Cognition* 3:25-62.

Damasio, A.R. 1990. Synchronous activation in multiple cortical regions: A mechanism for recall. *Seminars in the Neurosciences* 2:287-96.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., & Reitbock, H.J. 1988. Coherent oscillations: A mechanism for feature linking in the visual cortex. *Biological Cybernetics* 60:121-30.

Eckhorn, R., Reitbock, H.J., Arndt, M., & Dicke, P. 1989. A neural network for feature linking via synchronous activity: Results from cat visual cortex and from simulations. In (R. Cotterill, ed) *Models of Brain Function*. Cambridge University Press.

Engel, A.K., Konig, P. & Singer, W. 1991. Direct physiologic evidence for scene segmentation by temporal coding. *Proceedings of the National Academy of Sciences USA* 88:1936-40.

Engel, A.K., Konig, P., Kreiter, A.K. & Schillen, T.B. 1992. Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neurosciences* 15:218-26.

Engel, A.K., Fries, P., Konig, P., Brecht, M. & Singer, W. 1999. Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition* 8:128-51.

Engel, A.K., Fries, P., Konig, P., Brecht, M. & Singer, W. 1999. Does time help to understand consciousness? *Consciousness and Cognition* 8:260-68.

Fries, P., Roelfsema, P., Engel, A., & Singer, W. 1997. Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proceedings of the National Academy of Sciences USA* 94:12699-12704.

Gold, I. 1999. Does 40-Hz oscillation play a role in visual consciousness? *Consciousness and Cognition* 8:186-95.

Golledge, H.D.R., Hilgetag, C.C., & Tovee, M.J. 1996. Information processing: A solution to the binding problem. *Current Biology* 6:1092-95.

Gray, C.M., Konig, P., Engel, A.K. & Singer, W. 1992. Oscillatory responses in

cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338:334-7.

Gray, C.M. 1994. Synchronous oscillations in neuronal systems: Mechanisms and functions. *Journal of Computational Neuroscience* 1:11-38.

Hardcastle, V.G. 1994. Psychology's "binding problem" and possible neurobiological solutions. *Journal of Consciousness Studies* 1:66-90.

Hardcastle, V.G. 1996. How we get there from here: Dissolution of the binding problem. *Journal of Mind and Behavior* 17:251-66.

Hardcastle, V.G. 1997. Consciousness and the neurobiology of perceptual binding. *Seminars in Neurology* 17:163-70.

Konig, P. & Engel, A.K. 1995. Correlated firing in sensory-motor systems. *Current Opinion in Neurobiology* 5:511-19.

Konig, P., Engel, A.K. & Singer, W. 1995. Relation between oscillatory activity and long-range synchronization in cat visual cortex. *Proceedings of the National Academy of Sciences USA* 92:290-94.

Konig, P., Engel, A.K., Roelfsema, P.R. & Singer, W. 1995. How precise is neural synchronization? *Neural Computation* 7:469-85.

Llinas, R. & Ribary, U. 1998. Temporal conjunction in thalamocortical transactions. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Nelson, J.I. 1995. Binding in the visual system. In (M. Arbib, ed) *_Handbook of Brain Theory and Neural Networks_*. MIT Press.

Newman, J. & Grace, A.A. 1999. Binding across time: The selective gating of frontal and hippocampal systems modulating working memory and attentional states. *Consciousness and Cognition* 8:196-212.

Prinzmetal, W. 1981. Principles of feature integration in visual perception. *Perception and Psychophysics* 30:330-40.

Revonsuo, A. 1999. Binding and the phenomenal unity of consciousness. *Consciousness and Cognition* 8:173-85.

Sauve, K. 1999. Gamma-band synchronous oscillations: Recent evidence regarding their functional significance. *Consciousness and Cognition* 8:213-24.

Schillen, T.B. & Konig, P. 1994. Binding by temporal structure in multiple feature domains of an oscillatory neural network. *Biological Cybernetics* 5:397-405.

Shastri, L. & Ajjanagadde, V. 1993. From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic binding using temporal synchrony. *Behavioral and Brain Sciences* 16:417-51.

Sillito, A.M., Jones, H.E., Gerstein, G.L., & West, D.C. 1994. Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature* 369:479-82.

Singer, W. 1993. Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology* 55:349-74.

Singer, W. & Gray, C.M. 1995. Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* 18:555-86.

Singer, W., Engel, A.K., Kreiter, A., Munk, M., & Roelfsema, P. 1997. Neuronal assemblies: Necessity, signature, and detectability. *Trends in Cognitive Sciences* 1:252-60.

Steriade, M., McCormick, D.A., & Sejnowski, T.J. 1993. Thalamocortical oscillations in the sleeping and aroused brain. *Science* 262:679-85.

Steriade, M. 1998. Corticothalamic networks, oscillations, and plasticity. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Stryker, M. 1989. Is grandmother an oscillation? *Nature* 338:297-8.

Treisman, A. 1980. A feature integration theory of attention. *Cognitive Psychology* 12:97-136.

Treisman, A. 1996. The binding problem. *Current Opinion in Neurobiology* 6:171-8.

von der Malsburg, C. 1995. Binding in models of perception and brain function. *Current Opinion in Neurobiology* 5:520-28.

Usher, M. & Donnelly, N. 1998. Visual synchrony affects binding and segmentation in perception. *Nature* 394:179-82.

Wolfe, J.M. & Bennett, S.C. 1997. Preattentive object files: Shapeless bundles of basic features. *Vision Research* 37:25-43.

6.1h Consciousness and Anesthesia

Aitkenhead, A.R. 1993. Conscious awareness. In (P. Sebel, B. Bonke, & E. Winograd, eds) *_Memory and Awareness in Anesthesia_*. Prentice-Hall.

Alkire, M., Haier, R., & Fallon, J. 2000. Toward a unified theory of narcosis: Brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness. *Consciousness & Cognition* 9:370-386.

Andrade, J. 1995. Learning during anesthesia: A review. *British Journal of Psychology* 86:479-506.

Andrade, J. 1997. Investigations of hypesthesia: Using anesthetics to explore relationships between consciousness, learning, and memory. *Consciousness and Cognition* 5:562-80.

Andrade, J. & Jones, J.G. 1997. Awareness in anesthesia. In (G. Hall & M. Morgan, eds) *_Short Practice of Anesthesia_*. Chapman and Hall.

Andrade, J. 2000. NMDA receptor--mediated consciousness: A theoretical framework for understanding the effects of anesthesia on cognition? In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Bonke, B., Fitch, W. & Millar, K. (eds) 1990. *_Memory and Awareness In Anesthesia_*. Swets & Zeitlinger.

Bonke, B., Bovill, J.G., & Moerman, N. (eds) 1996. *_Memory and Awareness in Anesthesia III_*. Van Gorcum.

Caseley-Rondi, G., Merikle, P.M. & Bowers, K.S. 1994. Unconscious cognition in the context of general anesthesia. *Consciousness and Cognition* 3:166-95.

Cogliolo, P., Romano, V., Villani, R., & Galano, M. 1993. Effectiveness of Evans' technique for the evaluation of awareness. In (P. Sebel, B. Bonke, & E.

Winograd, eds) *_Memory and Awareness in Anesthesia 2_*. Prentice-Hall.

Corner, M. 1976. The nature of consciousness: some persistent conceptual difficulties and a practical suggestion. *Progress in Brain Research* 45:471-5.

Eich, E., Reeves, J.L., & Katz, R.L. 1985. Anesthesia, amnesia, and the memory/awareness distinction. *Anesthesia and Analgesia* 64:1143-48.

Evans, J.M. 1987. Patient's experiences of awareness during general anesthesia. In (M. Rosen & J. Lunn, eds) *_Consciousness, Awareness, and Pain in General Anesthesia_*. Butterworths.

Flohr, H. 1995. An information-processing theory of anesthesia. *Neuropsychologia* 33:1169-80.

Flohr, H. 1998. On the mechanism of action of anesthetic agents. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Flohr, H. 2000. NMDA-receptor-mediated computational processes and phenomenal consciousness. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Franks, N.P. & Lieb, W.R. 1998. The molecular basis of general anesthesia: Current ideas. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Franks, N. & Lieb, W. 2000. The role of NMDA receptors in consciousness: What we learn from anesthetic mechanisms? In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Ghoneim, M.M. & Block, R.I. 1992. Learning and consciousness during general anesthesia. *Anesthesiology* 76:279-305.

Hagan, S., Jibu, M. & Yasue, K. 1994. Consciousness and anesthesia: A hypothesis involving biophoton emission in the microtubular cytoskeleton of the brain. In (K. Pribram, ed) *_Origins: Brain and Self-organization_*. Lawrence Erlbaum.

Hameroff, S. 2001. Anesthesia: The "other side" of consciousness. *Consciousness and Cognition* 10:217-229.

Hill, D.S., & Hill, D.S. 1910. The loss and recovery of consciousness under anesthesia. *Psychological Bulletin* 7:77-83.

Jansen, C.K., Bonke, B., Klein, J. & Bezstarosti, J. 1990. Unconscious perception during balanced anesthesia? In (B. Bonke, W. Fitch, & K. Millar, 1990) *_Memory and Awareness in Anesthesia_*. Swets & Zeitlinger.

Jones, J.G. 1988. Awareness during anesthesia. *Anaesthesia Rounds*.

Kihlstrom, J.F. & Schacter, D.L. 1990. Anesthesia, amnesia, and the cognitive unconscious. In (B. Bonke, W. Fitch, & K. Millar, 1990) *_Memory and Awareness in Anesthesia_*. Swets & Zeitlinger.

Kihlstrom, J.F. & Couture, L.J. 1992. Awareness and information processing during general anesthesia. *Journal of Psychopharmacology* 6:410-17.

Kiviniemi K. 1994. Conscious awareness and memory during general anesthesia. *Aana Journal* 62:441-9.

Kulli, J. & Koch, C. 1991. Does anaesthesia cause loss of consciousness? *Trends in Neuroscience*, 14, 6-10.

Levinson, B.W. 1965. States of awareness during general anaesthesia. *British Journal of Anaesthesia* 37:544-546.

Lewis, S.A., Jenkinson, J. & Wilson, J. 1973. An EEG investigation of awareness during anaesthesia. *British Journal of Psychology* 64:413-5.

Merikle, P. M., & Daneman, M. 1996. Memory for unconsciously perceived events: Evidence from anesthetized patients. *Consciousness and Cognition* 5:525-541.

Merikle, P.M. & Daneman, M. 1996. Memory for events during anesthesia: A meta-analysis. In (B. Bonke, J.G. Bovill, & N. Moerman, eds) *Memory and Awareness in Anesthesia III*. Van Gorcum.

Moerman, N, Bonke, B. & Oosting, J. 1993. Awareness and recall during general anesthesia: Facts and feelings. *Anesthesiology* 79:454-64.

Mostert, J.W. 1975. States of awareness during general anesthesia. *Perspectives in Biology and Medicine* 19:68-76.

Munglani, R. & Jones, J.G. 1992. Sleep and general anesthesia as altered states of consciousness. *Journal of Psychopharmacology* 6:399-409.

Nikolinakos, D. 1994. General anesthesia, consciousness, and the skeptical challenge. *Journal of Philosophy* 2:88-104.

Plourde, G. 2001. Identifying the neural correlates of consciousness: Strategies with general anesthetics. *Consciousness and Cognition* 10:241-44.

Rosen, M., & Lunn, J.N. (eds) 1987. *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.

Sebel, P.S., Bonke, B. & Winograd, E. (eds) 1993. *Memory and Awareness in Anesthesia 2*. Prentice-Hall.

Tinnin, L. 1994. Conscious forgetting and subconscious remembering of pain. *Journal of Clinical Ethics* 5:151-52.

Tracy, J. 1993. Awareness in the operating room: A patient's view. In (P. Sebel, B. Bonke, & E. Winograd, eds) *Memory and Awareness in Anesthesia*. Prentice-Hall.

Utting, J.E. 1987. Awareness: Clinical aspects. In (M. Rosen & J. Lunn, eds) *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.

White, D.C. 1987. Anesthesia: A privation of the senses: An historical introduction and some definitions. In (M. Rosen & J. Lunn, eds) *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.

Vickers, M.D. 1987. Detecting consciousness by clinical means. In (M. Rosen & J. Lunn, eds) *Consciousness, Awareness, and Pain in General Anesthesia*. Butterworths.

6.1i Foundational Issues

Baars, B.J. 2001. The brain basis of a "consciousness monitor":Scientific and medical significance. *Consciousness and Cognition* 10:159-164.

Bisiach, E. 1988. The (haunted) brain and consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science*. Oxford University Press.

Block, N. 1998. How to find the neural correlate of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*

1996_. MIT Press.

Buck, R. 1993. What is this thing called subjective experience? Reflections on the neuropsychology of qualia. *Neuropsychology* 7:490-99.

Chalmers, D.J. 1998. On the search for the neural correlate of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Chalmers, D.J. 2000. What is a neural correlate of consciousness? In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Churchland, P.S. 1988. Reduction and the neurobiological basis of consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Churchland, P.S. 1994. Can neurobiology teach us anything about consciousness? *Proceedings and Addresses of the American Philosophical Association* 67:23-40.

Cleeremans, A. & Haynes, J. 1999. Correlating consciousness: A view from empirical science. *Revue Internationale de Philosophie* 3:387-420.

Cobb, S. 1952. On the nature and locus of mind. *Archives of Neurology and Psychiatry* 67:172-7.

Creutzfeld, O.D. 1987. Inevitable deadlocks of the brain-mind discussion. In (B. Gulyas, ed) *_The Brain-Mind Problem: Philosophical and Neurophysiological Approaches_*. Leuven University Press.

Farber, I.B. & Churchland, P.S. 1995. Consciousness and the neurosciences: Philosophical and theoretical issues. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.

Freeman, W.J. 1997. Three centuries of category errors in studies of the neural basis of consciousness and intentionality. *Neural Networks* 10:1175-83.

Gillett, G. 1988. Consciousness and brain function. *Philosophical Psychology* 1:325-39.

Gillett, G. 1995. Consciousness, thought, and neurological integrity. *Journal of Mind and Behavior* 16:215-33.

Gloor, P. 1986. Consciousness as a neurological concept in epileptology: a critical review. *Epilepsia* 27:S14-26.

Gordon, G., Maxwell, G. & Savodnik, I. (eds) 1976. *_Consciousness and the Brain: A Scientific and Philosophical Inquiry_*. Plenum.

Gray, J. 1998. Creeping up on the hard question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Hamanaka, T. 1997. The concept of consciousness in the history of neuropsychiatry. *History of Psychiatry* 8:361-373.

Hardcastle, V.G. 1995. *_Locating Consciousness_*. John Benjamins.

Hardcastle, V.G. 1996. Discovering the moment of consciousness? I: Bridging techniques at work, & II: An ERP analysis of priming using novel visual stimuli. *Philosophical Psychology* 9:149-96.

Hardcastle, V. 2000. How to understand the N in NCC. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Ivanitsky A.M. 1993. Consciousness: criteria and possible mechanisms. *International Journal of Psychophysiology* 14:179-87.

Kinsbourne, M. 1997. What qualifies a representation for a role in consciousness? In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Lahav, R. 1993. What neuropsychology tells us about consciousness. *Philosophy of Science* 60:67-85.

Lahav, R. 1997. The conscious and the nonconscious: Philosophical implications of neuropsychology. In (M. Carrier & P. Machamer, eds) *_Mindsapes: Philosophy, Science, and the Mind_*. Pittsburgh University Press.

Newton, N. 1991. Consciousness, qualia, and re-entrant signaling. *Behavior and Philosophy* 19:21.

Niedermeyer E. 1994. Consciousness: Function and definition. *Clinical Electroencephalography* 25:86-93.

Revonsuo, A. 1998. How to take consciousness seriously in cognitive neuroscience. *Communication and Cognition* 30:185-205.

Revonsuo, A. 2000. Prospects for a scientific research program on consciousness. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Schiller, F. 1952. Consciousness reconsidered. *Archives of Neurology and Psychiatry* 67:199-227.

6.1j Consciousness and Neuroscience, Misc

Arhem, P. 1996. Vertical information flow in the brain: on neuronal micro events and consciousness. *Biosystems* 38:191-98.

Bakhman, T. 2000. *_Microgenetic Approach to the Conscious Mind_*. John Benjamins.

Beck, H. 1976. Neuropsychological servosystems, consciousness, and the problem of embodiment. *Behavioral Science* 21:139-60.

Bernhaut, M., Gellhorn, E. & Rasmussen, A.T. 1953. Experimental contributions to the problem of consciousness. *Journal of Neurophysiology* 16:21-35.

Boitano, J. 1996. Edelmans's biological theory of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Borrett, D., Kelly, S., & Kwan, H. 2000. Phenomenology, dynamical neural networks and brain function. *Philosophical Psychology* 13:213-228.

Calvin, W. 1990. *_The Cerebral Symphony: Seashore Reflections on the Structure of Consciousness_*. Bantam.

Cowey, A. 1997. Current awareness: Spotlight on consciousness. *Developmental Medicine and Child Neurology* 39:54-62.

Creutzfeld, O.D. 1979. Neurophysiological mechanisms and consciousness. In *_Brain and Mind_* (Ciba Foundation Symposium 69). Elsevier.

Crick, F. & Koch, C. 1992. The problem of consciousness. *Scientific American* 267(3):152-60.

Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. Scribners.

Damasio, A. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.

Dehaene, S. & Naccache, L. 2001. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79:1-37.

Delacour, J. 1995. An introduction to the biology of consciousness. *Neuropsychologia* 33:1061-1074.

Delacour, J. 1997. Neurobiology of consciousness: An overview. *Behavioural Brain Research* 85:127-141.

Delafresnaye, J.F. (ed) 1954. *Brain Mechanisms and Consciousness*. Blackwell.

Desmedt J.E. & Tomberg, C. 1995. Consciousness. *Electroencephalography and Clinical Neurophysiology, Supplement* 44:227-34.

Donald, M. 1995. The neurobiology of human consciousness: An evolutionary approach. *Neuropsychologia* 33:1087-1102.

Donchin, E., McCarthy, G., Kutas, M. & Ritter, W. 1983. Event-related brain potentials in the study of consciousness. In (R. Davidson, S. Schwartz, & D. Shapiro, eds) *Consciousness and Self-Regulation*. Plenum Press.

Donnelly G.F. 1982. Consciousness: the brain and self-regulation modalities. *Topics in Clinical Nursing* 3:13-20.

Doty, R.W. 1975. Consciousness from neurons. *Acta Neurobiologiae Experimentalis* 35:791-804.

Eccles, J.C. (ed) 1966. *Brain and Conscious Experience*. Springer.

Eccles, J.C. 1974. Cerebral activity and consciousness. In (F. Ayala & T. Dobzhansky, eds) *Studies in the Philosophy of Biology*. University of California Press.

Eccles, J.C. 1987. The effect of silent thinking on the cerebral cortex. In (B. Gulyas, ed) *The Brain-Mind Problem: Philosophical and Neurophysiological Approaches*. Leuven University Press.

Edelman, G.M. 1989. *The Remembered Present: A Biological Theory of Consciousness*. Basic Books.

Edelman, G.M. 1992. *Bright Air, Brilliant Fire: On the Matter of the Mind*. Penguin.

Fessard, A.E. 1952. Mechanisms of nervous integration and conscious experience. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.

Frith, C.D. 1992. Consciousness, information processing, and the brain. *Journal of Psychopharmacology* 6:436-40.

Frith, C.D. 1996. The role of the prefrontal cortex in self-consciousness: the case of auditory hallucinations. *Philosophical Transactions of the Royal Society of London* B351:1505-12.

Gastaut, H. 1954. The brain stem and cerebral electrogenesis in relation to consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.

Gazzaniga, M. 1988. Brain modularity: Toward a philosophy of conscious experience. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Globus, G., Maxwell, G. & Savodnik, I. (eds) 1975. *_Consciousness and the Brain_*. Plenum Press.

Greenfield, S. 1995. *_Journey to the Centers of the Mind_*. W.H. Freeman.

Jasper, H. & Shagass, C. 1941. Conscious time judgments related to conditioned time intervals and voluntary control of the alpha rhythm. *Journal of Experimental Psychology* 28:503-508.

Kety, S.S. 1952. Consciousness and the metabolism of the brain. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the Third Conference_*. Josiah Macy Foundation.

Kinsbourne, M. 1995. Models of consciousness: Serial or parallel in the brain? In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.

Kokoszka, A. 1993. Information metabolism as a model of consciousness. *International Journal of Neuroscience* 68:165-77.

Lin, S., Tsai, Y., & Liou, C. 1993. Conscious mental tasks and their EEG signals. *Medical and Biological Engineering and Computing* 31:421-26.

Luria, A. 1976. The human brain and conscious activity. In Schwartz & Shapiro 1978.

Mitterauer B. 1998. An interdisciplinary approach towards a theory of consciousness. *Biosystems* 45:99-121.

Pare, D. & Llinas, R. 1995. Conscious and pre-conscious processes as seen from the standpoint of sleep-waking cycle neurophysiology. *Neuropsychologia* 33:1155-1168.

Penfield, W. 1975. *_The Mystery of the Mind_*. Princeton University Press.

Petty, P.G. 1998. Consciousness: A neurosurgical perspective. *Journal of Consciousness Studies* 5:86-96.

Picton, T.W. & Stuss, D.T. 1994. Neurobiology of conscious experience. *Current Opinion in Neurobiology* 4:256-65.

Pribram, K.H. 1990. Brain and consciousness: A wealth of data. In (E. John, ed), *_Machinery of the Mind: Data, Theory, and Speculations about Higher Brain Function_*. Boston: Birkhauser.

Ramachandran, V.S. & Hirstein, W. 1998. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies* 4:429-57.

Rose, S. 1973. *_The Conscious Brain_*. Paragon House.

Rudell, A.P. & Hua, J. 1996. The recognition potential and conscious awareness. *Electroencephalography and Clinical Neurophysiology* 98:309-18.

Salazar, A., Grafman J., Vance S., Dillon J.D., & Ludlow, C. 1986. Consciousness and amnesia after penetrating head injury: neurology and anatomy. *Neurology* 36:178-87.

Singer, W. 2000. Phenomenal awareness and consciousness from a neurobiological perspective. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT

Press.

Simonov, P.V. 1994. Consciousness and the brain. *Neuroscience and Behavioral Physiology* 24:234-38.

Travis F.T. & Orme-Johnson D.W. 1989. Field model of consciousness: EEG coherence changes as indicators of field effects. *International Journal of Neuroscience* 49:203-11.

Trevarthen, C. 1979. The tasks of consciousness: How could the brain do them? In *_Brain and Mind_* (Ciba Foundation Symposium 69). Elsevier.

Turner B.H. & Knapp M.E. 1995. Consciousness: A neurobiological approach. *Integrative Physiological and Behavioral Science* 30:151-6.

Walter, W.G. 1954. Theoretical properties of diffuse projection systems in relation to behaviour and consciousness. In (J. Delafresnaye, ed) *_Brain Mechanisms and Consciousness_*. Blackwell.

Webb, A.C. 1970. Consciousness and the cerebral cortex. *British Journal of Anaesthesia* 55:209-19.

Zeman, A.Z.J., Grayling, A.C. & Cowey, A. 1997. Contemporary theories of consciousness. *Journal of Neurology, Neurosurgery, and Psychiatry* 62:549-552.

6.2 Consciousness and Psychology

6.2a Cognitive Models of Consciousness

Allport, D.A. 1979. Conscious and unconscious cognition: A computational metaphor for the mechanism of attention and integration. In (L. Nilsson, ed) *_Perspectives on Memory Research_*.

Baars, B.J. 1983. Conscious contents provide the nervous system with coherent, global information. In (R. Davidson, G. Schwartz, & D. Shapiro, eds) *_Consciousness and Self-Regulation_*. Plenum.

Baars, B.J. 1988. *_A Cognitive Theory of Consciousness_*. Cambridge University Press.

Baars, B.J. 1993. How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity? In *_Experimental and Theoretical Studies of Consciousness_* (Ciba Foundation Symposium 174).

Baars, B.J., Fehling, M.R., LaPolla, M., & McGovern, K. 1997. Consciousness creates access: Conscious goal images recruit unconscious action routines, but goal competition serves to "liberate" such routines, causing predictable slips. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Baars, B.J. 1997. *_In the Theater of Consciousness: The Workspace of the Mind_*. Oxford University Press.

Baars, B.J. 1997. In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies* 4:292-309.

Baars, B.J. 1998. Metaphors of consciousness and attention in the brain. *Trends in Neurosciences* 21:58-62.

Baars, B.J. & McGovern, K. 1996. Cognitive views of consciousness: What are

the facts? How can we explain them? In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.

Bechtel, W. 1995. Consciousness: Perspectives from symbolic and connectionist AI. *Neuropsychologia*.

Brown, R.A. 1997. Consciousness in a self-learning, memory-controlled, compound machine. *Neural Networks* 10:1333-85.

Browne, C., Evans, R., Sales, N., & Aleksander, I.L. 1997. Consciousness and neural cognizers: A review of some recent approaches. *Neural Networks* 10:1303-1316.

Burks, A.W. 1986. An architectural theory of functional consciousness. In (N. Rescher, ed) *_Current Issues in Teleology_*. University Press of America.

Cabanac, M. 1996. On the origin of consciousness, a postulate, and its corollary. *Neuroscience and Biobehavioral Reviews* 20:33-40.

Cam, P. 1989. Notes toward a faculty theory of cognitive consciousness. In (P. Slezak, ed) *_Computers, Brains and Minds_*. Kluwer.

Carr, T.H. 1979. Consciousness in models of human information processing: Primary memory, executive control, and input regulation. In (G. Underwood & R. Stevens, eds) *_Aspects of Consciousness, Volume 1_*. Academic Press.

Cotterill, R.M.J. 1996. Prediction and internal feedback in conscious perception. *Journal of Consciousness Studies* 3:245-66.

Cotterill, R.M.J. 1997. On the mechanism of consciousness. *Journal of Consciousness Studies* 4:231-48.

Cotterill, R.M.J. 1997. Navigation, consciousness and the body/mind "problem". *Psyke and Logos*, 18:337-341.

Farrell, B.A. 1970. The design of a conscious device. *Mind* 79:321-46.

Gregory R.L. 1984. Is consciousness sensational inferences? *Perception* 13:641-6.

Hardcastle, V.G. 1995. A critique of information processing theories of consciousness. *Minds and Machines* 5:89-107.

Harnad, S. 1982. Consciousness: An afterthought. *Cognition and Brain Theory* 5:29-47.

Harth, E. 1993. *_The Creative Loop: How the Brain Makes a Mind_*. Addison Wesley.

Harth, E. 1995. The sketchpad model: A theory of consciousness, perception, and imagery. *Consciousness and Cognition* 4:346-68.

Harth, E. 1996. Self-referent mechanisms as the neuronal basis of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Jackendoff, R. 1987. *_Consciousness and the Computational Mind_*. MIT Press.

John, E.R. 1976. A model of consciousness. In (G. Schwartz & D. Shapiro, eds) *_Consciousness and Self-Regulation_*. Plenum Press.

Johnson, M.K. & Reeder, J.A. 1997. Consciousness as meta-processing. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

- Johnson-Laird, P. 1983. A computational analysis of consciousness. *Cognition and Brain Theory* 6:499-508. Also in (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.
- Kawato, M. 1997. Bidirectional theory approach to consciousness. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *_Cognition, Computation, and Consciousness_*. Oxford University Press.
- Lauro-Grotto, R., Reich, S. & Virasoro, M.A. 1997. The computational role of conscious processing in a model of semantic memory. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *_Cognition, Computation, and Consciousness_*. Oxford University Press.
- Lloyd, D. 1995. Consciousness: A connectionist manifesto. *Minds and Machines* 5:161-85.
- Lloyd, D. 1996. Consciousness, connectionism, and cognitive neuroscience: A meeting of the minds. *Philosophical Psychology* 9:61-78.
- Mathis, D.W. & Moxer, M. 1995. On the computational utility of consciousness. In (G. Tesauro, D. Touretzky, & T. Leen, eds) *_Advances in Neural Information Processing Systems 7_*. MIT Press.
- Michie, D. 1994. Consciousness as an engineering issue (Parts 1 and 2). *Journal of Consciousness Studies* 1:192-95, 2:52-66.
- Norretranders, T. 1991. *_The User Illusion: Cutting Consciousness Down to Size_*. Viking Penguin.
- O'Brien, G. & Opie, J. 1998. A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences* 22:127-48.
- Oatley, K. 1981. Representing ourselves: Mental schemata, computational metaphors, and the nature of consciousness. In (G. Underwood & R. Stevens, eds) *_Aspects of Consciousness, Volume 2_*. Academic Press.
- Parsons, T. 1953. Consciousness and symbolic processes. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the Fourth Conference_*. Josiah Macy Foundation.
- Phaf, R.H. & Wolters, G. 1997. A constructivist and connectionist view on conscious and nonconscious processes. *Philosophical Psychology* 10:287-307.
- Restian, A. 1981. Informational analysis of consciousness. *International Journal of Neuroscience* 13:229-37.
- Revonsuo, A. 1993. Cognitive models of consciousness. In (M. Kamppinen, ed) *_Consciousness, Cognitive Schemata, and Relativism_*. Kluwer.
- Roberts, H. 1968. Consciousness in animals and automata. *Psychological Reports* 22:1226-28.
- Rolls, E.T. 1997. Consciousness in neural networks? *Neural Networks* 10:1227-1303.
- Schacter, D.L. 1989. On the relation between memory and consciousness: Dissociable interactions and conscious experience. In (H. Roediger & F. Craik, eds) *_Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving_*.
- Schneider, W. & Pimm-Smith, M. 1997. Consciousness as a message-aware control mechanism to modulate cognitive processing. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Shallice, T. 1972. Dual functions of consciousness. *Psychological Review* 79:383-93.

Shallice, T. 1978. The dominant action system: An information-processing approach to consciousness. In (K.S. Pope & J.L. Singer, eds) *_The Stream of Consciousness: Scientific Investigation into the Flow of Experience_*. Plenum.

Shallice, T. 1988. Information-processing models of consciousness: possibilities and problems. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.

Sommerhoff G. & MacDorman K. 1994. An account of consciousness in physical and functional terms: A target for research in the neurosciences. *Integrative Physiological and Behavioral Science* 29:151-81.

Sommerhoff, G. 1996. Consciousness as an internal integrating system. *Journal of Consciousness Studies* 3:139-57.

Strehler, B.L. 1989. Monitors: key mechanisms and roles in the development and aging of the consciousness and self. *Mechanisms of Ageing and Development* 47:85-132.

Sun, R. 1997. Learning, action, and consciousness: A hybrid approach toward modeling consciousness. *Neural Networks* 10:1317-33.

Sviderskaya, N.E. 1991. Consciousness and information selection. *Neuroscience and Behavioral Physiology* 21:526-31.

Taylor, J.G. 1996. Modeling what it is like to be. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Taylor, J.G. 1996. A competition for consciousness? *Neurocomputing* 11:271-96.

Taylor, J.G. 1997. Neural networks for consciousness. *Neural Networks* 10:1207-27.

Taylor, J.G. & Mueller-Gaertner, H. 1997. Non-invasive analysis of awareness. *Neural Networks* 10:1185-1194.

Werbos, P. 1997. Optimization: A foundation for understanding consciousness. In (D. Levine & W. Elsberry, eds) *_Optimality in Biological and Artificial Networks?_*. Lawrence Erlbaum.

6.2b Unconscious Perception

Adams, J.K. 1957. Laboratory studies of behavior without awareness. *Psychological Bulletin* 54:383-405.

Balay, J. & Shevrin, H. 1988. The subliminal psychodynamic activation method: A critical review. *American Psychologist* 43:161-74.

Bar, M. 2000. Conscious and nonconscious processing of visual object identity. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

Bargh, J.A. 1992. Does subliminality matter to social psychology? Awareness of the stimulus versus awareness of its influence. In (R. Bornstein & T. Pittman, eds) *_Perception without Awareness_*. Guilford.

Bhalla, M. & Proffitt, D. 2000. Geographical slant perception: Dissociation and coordination between explicit awareness and visually guided actions. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

- Bornstein, R.F. 1989. Exposure and affect: Overview and meta-analysis of research 1968-1987. *Psychological Bulletin* 106:265-89.
- Bornstein, R.F. 1992. Subliminal mere exposure effects. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.
- Bornstein, R.F. & Pittman, T.S. 1992. *Perception without Awareness: Cognitive, Clinical, and Social Perspectives*. Guilford.
- Bowers, K.S. 1982. On being unconsciously influenced and informed. In (K. Bowers & D. Meichenbaum, eds) *The Unconscious Reconsidered*. Wiley.
- Cheesman, J. & Merikle, P.M. 1984. Priming with and without awareness. *Perception and Psychophysics* 36:387-95.
- Cheesman, J. & Merikle, P.M. 1986. Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology* 40:343-67.
- Damian, M. 2001. Congruity effects evoked by subliminally presented primes: Automaticity rather than semantic processing. *Journal of Experimental Psychology: Human Perception & Performance* 27:154-165.
- Debner, J.A. & Jacoby, L.L. 1994. Unconscious perception: Attention, awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20:304-17.
- Dixon, N.F. 1971. *Subliminal Perception: The Nature of a Controversy*. McGraw-Hill.
- Dixon, N.F. & Henley, S.H.A. 1980. Without awareness. In (M. Jeeves, ed) *Psychology Survey 3*. Allen and Unwin.
- Doyle, J.R. 1990. Detectionless processing with semantic activation? A footnote to Greenwald, Klinger, and Liu (1989). *Memory and Cognition* 18:428-9.
- Eagle, M. 1959. The effects of subliminal stimuli of aggressive content upon conscious cognition. *Journal of Personality* 27:578-600.
- Erdelyi, M.H. 1970. Recovery of unavailable perceptual input. *Cognitive Psychology* 1:99-113.
- Eriksen, C.W. 1956. An experimental analysis of subception. *American Journal of Psychology* 69:625-34.
- Eriksen, C.W. 1956. Subception: Fact or artifact? *Psychological Review* 63:74-80.
- Eriksen, C.W. 1960. Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review* 67:279-300.
- Farah, M.J., Monheit, M.A. & Wallace, M.A. 1991. Unconscious perception of "extinguished" visual stimuli: Reassessing the evidence. *Neuropsychologia* 29:949-58.
- Fowler, C.A., Woldford, G., Slade, R. & Tassinary, L. 1981. Lexical access with and without awareness. *Journal of Experimental Psychology: General* 110:341-62.
- Fuhrer, M.J. & Eriksen, C.W. 1960. The unconscious perception of the meaning of verbal stimuli. *Journal of Abnormal and Social Psychology* 61:432-9.
- Goldiamond, I. 1958. Indicators of perception: 1. Subliminal perception, subception, unconscious perception: An analysis in terms of psychophysical

indicator methodology. *Psychological Bulletin* 55:373-411.

Greenwald, A.G., Klinger, M.R., & Liu, T.J. 1989. Unconscious processing of dichoptically masked words. *Memory and Cognition* 17:35-47.

Greenwald, A.G., Spangenberg, E., Pratkanis, A.R., & Eskenazi, J. 1991. Double blind tests of subliminal self-help audiotapes. *Psychological Science* 2:119-22.

Greenwald, A.G., Klinger, M.R. & Schuh, E.S. 1995. Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General* 124:22-42.

Greenwald, A.G. & Draine, S. 1997. Do subliminal stimuli enter the mind unnoticed?: Tests with a new method. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Hardaway, R.A. 1990. Subliminally activated symbiotic fantasies: Facts and artifacts. *Psychological Bulletin* 107:177-95.

Henley, S.H. 1984. Unconscious perception revisited: A comment on Merikle (1992). *Bulletin of the Psychonomic Society* 22:121-4.

Holender, D. 1986. Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences* 9:1-23.

Kemp-Wheeler, S.M. & Hill, A.B. 1988. Semantic priming without awareness: Some methodological considerations and implications. *Quarterly Journal of Experimental Psychology* 40:671-92.

Khurana, B. 2000. Face representation without conscious processing. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Kihlstrom, J.F. 1996. Perception without awareness of what is perceived, learning without awareness of what is learned. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Kihlstrom, J.F., Barnhardt, T.M. & Tataryn, D.J. 1992. Implicit perception. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.

Klauer, K. & Greenwald, A. 2000. Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception & Performance* 26:1506-1508.

Kostandov, E.A. 1994. Subsensory reactions and the problem of unconscious perception. *Sensory Systems* 7:149-53.

Krosnick, J.A., Betz, A.L., Jussim, L.J. & Lynn, A.R. 1992. Subliminal conditioning of attitudes. *Personality and Social Psychology Bulletin* 18:152-62.

Kunst-Wilson, W.R. & Zajonc, R.B. 1980. Affective discrimination of stimuli that cannot be recognized. *Science* 207:557-58.

Kunzendorf, R.G. 1985. Subconscious percepts as "unmonitored" percepts: An empirical study. *Imagination, Cognition and Personality* 4:365-73.

Lazarus, R.S. & McCleary R.A. 1983. Autonomic discrimination without awareness: A study of subception. *Psychological Review* 58:113-22.

Lewicki, P., Hill, T. & Czyewska, M. 1992. Nonconscious acquisition of information. *American Psychologist* 47:792-801.

- Lewis, J.L. 1970. Semantic processing of unattended messages using dichotic listening.
- MacLeod, C. 1998. Implicit perception: Perceptual processing without awareness. In (K. Kirsner, G. Speelman, eds) *_Implicit and Explicit Mental Processes_*. Lawrence Erlbaum.
- Marcel, A.J. 1983. Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology* 15:197-237.
- Marcel, A.J. 1983. Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology* 15:238-300.
- Merikle, P.M. 1982. Unconscious perception revisited. *Perception and Psychophysics* 31:298-301.
- Merikle, P.M. & Reingold, E.M. 1990. Recognition and lexical decision without detection: Unconscious perception? *Journal of Experimental Psychology: Human Perception and Performance* 16:574-83.
- Merikle, P.M. 1992. Perception without awareness: Critical issues. *American Psychologist* 47:792-5.
- Merikle, P.M. & Reingold, E.M. 1992. Measuring unconscious processes. In (R. Bornstein & T. Pittman, eds) *_Perception without Awareness_*. Guilford.
- Merikle, P.M., Joordens, S. & Stolz, J.A. 1995. Measuring the relative magnitude of unconscious influences. *Consciousness and Cognition* 4:422-39.
- Merikle, P.M. & Daneman, M. 1997. Psychological investigations of unconscious perception. *Journal of Consciousness Studies*.
- Merikle, P. & Daneman, M. 2000. Conscious vs. unconscious perception. In (M. Gazzaniga, ed) *_The New Cognitive Neurosciences: 2nd Edition_*. MIT Press.
- Merikle, P.M., D. Smilek, & J.D. Eastwood 2001. Perception without awareness: Perspectives from cognitive psychology. *Cognition* 79:115-34.
- Miller, J. 2000. Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception & Performance* 26:1461-1477.
- Moore, T.E. 1992. Subliminal perception: Facts and fallacies. *Skeptical Inquirer* 16:273-81.
- Neuberg, S.L. 1988. Behavioral implications of information presented outside of conscious awareness: The effect of subliminal presentation of trait information on behavior in the Prisoner's Dilemma game. *Social Cognition* 6:207-30.
- Nolan, K.A. & Caramazza, A. 1982. Unconscious perception of meaning: A failure to replicate. *Bulletin of the Psychonomic Society* 20:23-26.
- Peirce, C.S. & Jastrow, J. 1884. On small differences in sensation. *Memoirs of the National Academy of Sciences* 3:75-83.
- Pisella, L. & Rosetti, Y. 2000. Interaction between conscious identification and non-conscious sensory-motor processing: Temporal constraints. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.
- Poppel, E., Held, R. & Frost, D. 1973. Residual function after brain wounds involving the central visual pathways in man. *Nature* 243:295-96.

Pratkanis, A.R. & Greenwald, A.G. 1988. Recent perspectives on unconscious processing: Still no marketing applications. *Psychology and Marketing* 5:337-53.

Purcell, D.G., Stewart, A.L. & Stanovich, K.K. 1983. Another look at semantic priming without awareness. *Perception and Psychophysics* 34:65-71.

Reingold, E.M. & Merikle, P.M. 1988. Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics* 44:563-575.

Reingold, E. & Merikle, P. 1991. Theory and measurement in the study of unconscious processes. *Mind and Language* 5:9-28.

Schmidt, T. 2000. Visual perception without awareness: priming responses by color. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Shevrin, H. 1992. Unconscious perception, memory, and consciousness: Cognitive and dynamic perspectives. In (R. Bornstein & T. Pittman, eds) *_Perception without Awareness_*. Guilford.

Silverman, L.H. & Weinberger, J. 1985. Mommy and I are one: Implications for psychotherapy. *American Psychologist* 40:1296-1308.

Theus, K.T. 1994. Subliminal advertising and the psychology of processing unconscious stimuli: A review of research. *Psychology and Marketing* 11:271-290.

Young, A. & Ellis, H. 2000. Overt and covert face recognition. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.

Weinberger, J. & Hardaway, R. 1990. Separating science from myth in subliminal psychodynamic activation. *Clinical Psychological Review* 10:727-56.

6.2c Unconscious Processes

Abrams, R. & Greenwald, A. 2000. Parts outweigh the whole (word) in unconscious analysis of meaning. *Psychological Science* 11:118-124.

Bridgeman, B. 1992. Conscious vs unconscious processes: The case of vision. *Theory and Psychology* 2:73-88.

Buchner, A. 1997. Consciousness, intention, and the process dissociation procedure. *Sprache and Kognition* 16:176-182.

Cowan, N., Stadler, M.A. 1996. Estimating unconscious processes: Implications of a general class of models. *Journal of Experimental Psychology: General* 125:195-200.

Dixon, N.F. 1981. *_Preconscious Processing_*. Wiley.

Erdelyi, M.H. 1974. A new look at the New Look: Perceptual defense and vigilance. *Psychological Review* 81:1-25.

Erdelyi, M.H. 1992. Psychodynamics and the unconscious. *American Psychologist* 47:784-87.

Field, A. 2000. I like it, but I'm not sure why: Can evaluative conditioning occur without conscious awareness? *Consciousness & Cognition* 9:13-36.

Ford, T. & Thompson, E. 2000. Preconscious and postconscious processes underlying construct accessibility effects: An extended search model.

Personality & Social Psychology Review 4:317-336.

Gaito, J. 1964. Stages of perception, unconscious processes, and information extraction. Journal of General Psychology 70:183-197.

Greenwald, A.G. 1992. New Look 3: Unconscious cognition reclaimed. American Psychologist 47:766-79.

Hilgard, E.R. 1958. Unconscious Processes and Man's Rationality. University of Illinois Press.

Hoffman, R. 1997. What neural network studies suggest regarding the boundary between conscious and unconscious mental processes. In (D. Stein, ed) Cognitive Science and the Unconscious. American Psychiatric Press

Hommel, B. 2000. Intentional control of automatic stimulus-response translation. In (Y. Rossetti & A. Revonsuo, eds) Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing. John Benjamins.

Imanaka, K. & Abernethy, B. 2000. Distance-location interference in movement reproduction: An interaction between conscious and unconscious processing? In (Y. Rossetti & A. Revonsuo, eds) Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing. John Benjamins.

Jacoby, L.L., Toth, J.P., Lindsay, D..S., Debner, J.A. 1992. Lectures for a layperson: Methods for revealing unconscious processes. In (R. Bornstein & B. Pittman, eds) Perception without Awareness: Cognitive, Clinical, and Social Perspectives. Guilford Press.

Kihlstrom, J.F. 1984. Conscious, subconscious, unconscious: A cognitive perspective. In (K.S. Bowers & D. Meichenbaum, eds) The Unconscious Reconsidered. Wiley.

Kihlstrom, J.F. 1987. The cognitive unconscious. Science 237:1445-1452.

Kihlstrom, J.F. 1990. The psychological unconscious. In (L. Pervin, ed) Handbook of Personality: Theory and Research. Guilford Press.

Kihlstrom, J.F., Barnhardt, T.M. & Tattryn, D.J. 1992. The psychological unconscious: Found, lost, and regained. American Psychologist 47:788-91.

Kihlstrom, J.F. 1995. The rediscovery of the unconscious mind. In (H. Morowitz & J. Singer, eds) The Mind, the Brain, and Complex Adaptive Systems. Addison-Wesley.

Kihlstrom, J.F. 1996. Unconscious processes in social interaction. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness. MIT Press.

Klinger, M., Burton, P., & Pitts, G. 2000. Mechanisms of unconscious priming: Response competition, not spreading activation. Journal of Experimental Psychology: Learning, Memory, & Cognition 26:441-455.

Lewicki, P. 1986. Nonconscious Social Information Processing. Academic Press.

Lewicki, P. & Hill, T. 1987. Unconscious processes as explanations of behavior in cognitive, personality, and social psychology. Personality and Social Psychology Bulletin 13:355-362.

Loftus, E.F. & Klinger, M.R. 1992. Is the unconscious smart or dumb? American Psychologist 47:761-65.

- Miller, J.G. 1951. Unconscious processes and perception. In (R. Blake & G. Ramsey, eds) *_Perception_*.
- Miller, J.G. 1952. The experimental study of unconscious processes. In (M. Reymert, ed) *_Feelings and Emotions_*.
- Perner, J. & Clements, W. 2000. From an implicit to an explicit "theory of mind". In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.
- Peterfreund, E. & Schwartz, J.T. 1971. Information processing and the nature of conscious and unconscious processes. *Psychological Issues* 7:219-29.
- Posner, M.I. 1991. Recent experimental studies of conscious and unconscious processes. In (M. Posner, B. Dwivedi, & I. Singh, eds) *_Contemporary Approaches to Cognitive Psychology_*. Rishi Publications.
- Radil, T., Radilova, J., Bozkov, V., & Bohdanecky, Z. 1981. Unconscious and conscious processes during visual perception. *Acta Neurobiologiae Experimentalis* 41:565-572.
- Revonsuo, A., Johanson, M., Wedlund, J., & Chaplin, J. 2000. The zombies among us: Consciousness and automatic behaviour. In (Y. Rossetti & A. Revonsuo, eds) *_Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing_*. John Benjamins.
- Schacter, D.L. 1992. Implicit knowledge: New perspectives on unconscious processes. *Proceedings of the National Academy of Sciences USA* 89:11113-17.
- Schwartz, M. 1981. Criteria for physiological substrates of unconscious processes. *American Psychologist* 36:434-435.
- Shevrin, H. & Dickman, S. 1980. The psychological unconscious: A necessary assumption for all psychological theory? *American Psychologist* 35:421-34.
- Shevrin, H., Smith, W.H., Fitzler, D.E. 1971. Average evoked response and verbal correlates of unconscious mental processes. *Psychophysiology* 8:149-62.
- Shevrin, H. & Fritzler, D.E. 1968. Visual evoked response correlates of unconscious mental processes. *Science* 161:295-298.
- Spitz, H.H. 1993. The role of the unconscious in thinking and problem solving. *Educational Psychology* 13:229-244.
- Spitz, H.H. 1995. Calendar calculating idiots savants and the smart unconscious. *New Ideas in Psychology* 13:167-182.
- Stein, D.J. (ed) 1997. *_Cognitive Science and the Unconscious_*. American Psychiatric Press.
- Suzuki, K. & Yamadori, A. 2000. Intact verbal description of letters with diminished awareness of their forms. *Journal of Neurology, Neurosurgery & Psychiatry* 68:782-786.
- Underwood, G. & Bright, J.E.H. 1996. Cognition with and without awareness. In (G. Underwood, ed) *_Implicit Cognition_*. Oxford University Press.
- Wachtel, P.L. 1987. "The unconscious" and unconscious processes. *Canadian Psychology* 28:107-108.
- Yu, J. & Bellezza, F. 2000. Process dissociation as source monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 26:1518-1533.

6.2d Visual Consciousness [see also 6.1b, 6.1c, 6.2b, 6.2e, 6.2j]

Bachmann, T. 1997. Visibility of brief images: The dual-process approach. *Consciousness and Cognition* 6:491-518.

Bachmann, T. 1998. Fast dynamics of visibility of brief visual images: The perceptual-retouch viewpoint. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Baxt, N. 1982. On the time necessary for a visual impression to come into consciousness. *Psychological Research* 44:1-12.

Blackmore, S.J., Brelstaff, G., Nelson, K. & Troscianko, T. 1995. Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. *Perception* 24:1075-81.

Bridgeman, B., Hendry, D. & Stark, L. 1975. Failure to detect displacements of the visual world during saccadic eye movements. *Vision Research* 15:719-22.

Bridgeman, B. 2000. Interactions between vision for perception and vision for behavior. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Cogan, A.I. 1995. Vision comes to mind. *Perception* 24:811-26.

Di Lollo, V., Enns, J., & Rensink, R. 2000. Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal Of Experimental Psychology-General* 129:481-507.

Dixon, M., Smilek, D., Cudahy, C., & Merikle, P. 2000. Five plus two equals yellow: Mental arithmetic in people with synaesthesia is not coloured by visual experience. *Nature* 406:365.

Durgin, F.H. 1995. On the filling in of the visual blind spot: Some rules of thumb. *Perception* 24:827-40.

Fernandez-Duque, D. & Thornton, I. 2000. Change detection without awareness: Do explicit reports underestimate the representation of change in the visual system? *Visual Cognition* 7:323-344.

Grimes, J. 1996. On the failure to detect changes in scenes across saccades. In (K. Akins, ed) *Perception*. Oxford University Press.

Irwin, D.E. 1991. Information integration across saccadic eye movements. *Cognitive Psychology* 23:420-56.

Iwasaki, S. 1993. Spatial attention and two modes of visual consciousness. *Cognition* 49:211-233.

Lachter, J., Durgin, F., & Washington, T. 2000. Disappearing percepts: Evidence for retention failure in metacontrast masking. *Visual Cognition* 7:269-279.

Mack, A. & Rock, I. 1998. *Inattentional Blindness*. MIT Press.

McConkie, G.W. & Zola, D. 1979. Is visual information integrated across successive fixations in reading? *Perception and Psychophysics* 25:221-24.

Natsoulas, T. 1993. An introduction to reflective seeing. *Journal of Mind and Behavior* 14:235-56.

Natsoulas, T. 1994. An introduction to reflective seeing. *Journal of Mind and Behavior* 15:351-74.

Nijhawan, R. & Khurana, B. 2000. Conscious registration of continuous and discrete visual events. In (T. Metzinger, ed) *_Neural Correlates of Consciousness_*. MIT Press.

Noe, A., Pessoa, L., & Thompson, E. 2000. Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition* 7: 93-106.

O'Regan, J.K. 1992. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology* 46:461-88.

O'Regan, J., Deubel, H., Clark, J., & Rensink, R. 2000. Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition* 7:191-211.

Palmer, S. 1999. *_Vision Science: Photons to Phenomenology_*. MIT Press.

Pashler, H. 1988. Familiarity and visual change detection. *Perception and Psychophysics* 41:191-201.

Pessoa, L., Thompson, E. & Noe, A. 1998. Finding out about filling in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*.

Phillips, W.A. 1974. On the distinction between sensory storage and visual short-term memory. *Perception and Psychophysics* 16:283-90.

Ramachandran, V.S. 1992. Filling in gaps in perception: Part I. *Current Directions in Psychological Science* 1:199-205.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8:368-373.

Rensink, R., O'Regan, J., & Clark, J. 2000. On the failure to detect changes in scenes across brief interruptions. *Visual Cognition Special* 7:127-145.

Scott-Brown, K., Baker, M., & Orbach, H. 2000. Comparison blindness. *Visual Cognition* 7:253-267.

Shapiro, K. 2000. Change blindness: Theory or paradigm? *Visual Cognition* 7:83-91.

Shore, D., & Klein, R. 2000. The effects of scene inversion on change blindness. *Journal of General Psychology* 127:27-43.

Simons D.J. & Levin, D.T. 1997. Change blindness. *Trends in Cognitive Science* 1:241-82.

Simons, D. 2000. Current approaches to change blindness. *Visual Cognition* 7:1-15.

Sperling, G. 1960. The information available in visual presentations. *Psychological Monographs* 74:1-29.

Wallis, G. & Buelthoff, H. 2000. What's scene and not seen: Influences of movement and task upon what we see. *Visual Cognition* 7:175-190.

6.2e Consciousness and Attention [see also 6.2d]

Arvidson, P.S. 1996. Toward a phenomenology of attention. *Human Studies* 19:71-84.

Baars, B.J. 1997. Some essential differences between consciousness and

attention, perception, and working memory. *Consciousness and Cognition* 6:363-371.

Baars, B.J. 1999. Attention vs. consciousness in the visual brain: Differences in conception, phenomenology, behavior, neuroanatomy, and physiology. *Journal of General Psychology* 126:224-33.

Baddeley, A.D. & Weiskrantz, L. (eds) 1993. *Attention: Selection, Awareness, and Control*. Oxford University Press.

Bridgeman, B. 1986. Relations between the physiology of attention and the physiology of consciousness. *Psychological Research* 48:259-266.

Cobb, C. 1955. Awareness, attention, and physiology of the brain stem. In (P. Hoch & J. Zubin, eds) *Experimental Psychopathology*.

Coslett, H.B. 1997. Consciousness and attention. *Seminars in Neurology* 17:137-44.

Cowan, N. & Wood, N.L. 1997. Constraints on awareness, attention, processing, and memory: Some recent investigations with ignored speech. *Consciousness and Cognition* 6:182-203.

Csikszentmihalyi, M. 1978. Attention and the holistic approach to behavior. In (K. Pope & J. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Hardcastle, V.G. 1997. Attention versus consciousness: A distinction with a difference. Manuscript.

He, S., Cavanagh, P. & Intriligator, J. 1996. Attentional resolution and the locus of visual awareness. *Nature* 383:334-37.

Hochberg, J. 1970. Attention, organization, and consciousness. In (D. Mostofsky, ed) *Attention: Contemporary Theory and Analysis*. Appleton-Century-Crofts.

Iwasaki, S. 1993. Spatial attention and two modes of visual consciousness. *Cognition* 49:211-233.

LaBerge, D. 1997. Attention, awareness, and the triangular circuit. *Consciousness and Cognition* 9:149-81.

LaBerge, D., Auclair, L., & Sieroff, E. 2000. Preparatory attention: Experiment and theory. *Consciousness & Cognition* 9:396-434.

Loper, A.B. & Hallahan, D.P. 1982. Meta-attention: The development of awareness of the attentional process. *Journal of General Psychology* 106:27-33.

Mack, A. & Rock, I. 1998. *Inattentional Blindness*. MIT Press.

Merikle, P.M. & Joordens, S. 1997. Parallels between perception without attention and perception without awareness. *Consciousness and Cognition* 6:219-36.

McCormick, P.A. 1997. Orienting attention without awareness. *Journal of Experimental Psychology: Human Perception & Performance* 23:168-180.

Newman, J.B., Baars, B.J., & Cho, S. 1997. A neural global workspace model for conscious attention. *Neural Networks* 10:1195-1206.

Newman, J.B. 1995. Thalamic contributions to attention and consciousness. *Consciousness and Cognition* 4:172-93.

Newsome, W.T. 1996. Visual attention: spotlights, highlights and visual awareness. *Current Biology* 6:357-60.

Posner, M.I. 1994. Attention: The mechanisms of consciousness. *Proceedings of the National Academy of Sciences USA* 91:7398-7403.

Posner, M.I. & Rothbart, M.K. 1992. Attentional mechanisms and conscious experience. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Prinzmetal, W. Amiri, H., Allen, K. & Edwards, T. 1997. The phenomenology of attention, part 1: Color, location, orientation, and "clarity". *Journal of Experimental Psychology: Human Perception and Performance*.

Prinzmetal, W., Nwachuku, I., Bodanski, L., & Blumenfeld, L. 1997. The phenomenology of attention, part 2: Brightness and contrast. *Consciousness and Cognition* 6:372-412.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8:368-373.

Rensink, R. 2000. Visual search for change: A probe into the nature of attentional processing. *Visual Cognition* 7:345-376.

Scheier, M.F. Matthews, K.A. & Carver, C.S. 1983. Focus of attention and awareness of bodily states. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Scholl, B. 2000. Attenuated change blindness for exogenously attended items in a flicker paradigm. *Visual Cognition* 7:377-396.

Shiffrin, R.M. 1997. Attention, automatism, and consciousness. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Smilek, D., Eastwood, J., & Merikle, P. 2000. Does unattended information facilitate change detection? *Journal of Experimental Psychology: Human Perception & Performance* 26:480-487.

Umla, C. & Moscovitch, M. 1994. *Attention and Performance 15: Conscious and Nonconscious Information Processing*. MIT Press.

Underwood, G. 1977. Attention, awareness, and hemispheric differences in word recognition. *Neuropsychologia* 15:61-67.

Underwood, G. 1983. Selective attention and selective awareness of conscious processes. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

6.2f Consciousness and Memory

Alkire, M.T., Haier, R.J., Fallon, J.H., & Barker, S.J. 1996. PET imaging of conscious and unconscious verbal memory. *Journal of Consciousness Studies* 3:448-62.

Allik, J. 2000. Available and accessible information in memory and vision. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Andreasen, N. 2000. Is schizophrenia a disorder of memory or consciousness? In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Baddeley, A. 1993. Working memory and conscious awareness. In (A. Collins, S. Gathercole, M. Conway, & P. Morris, eds) *Theories of Memory*. Lawrence Erlbaum.

Balint, E. 1987. Memory and consciousness. *International Journal of Psychoanalysis* 68:475-483.

Barba, G. 2000. Memory, consciousness, and temporality: What is retrieved and who exactly is controlling the retrieval? In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Brainerd, C.J., Stein, L.M., & Reyna, V.F. 1998. On the development of conscious and unconscious memory. *Developmental Psychology* 34:342-357.

Brewer, W.F. 1992. Phenomenal experience in laboratory and autobiographical memory. In (M. Conway, D. Rubin, H. Spinnler, & W. Wagenaar, eds) *Theoretical Perspectives on Autobiographical Memory*. Kluwer.

Brewer, W.F. 1996. What is recollective memory? In (D. Rubin, ed) *Remembering our Past: Studies in Autobiographical Memory*. Cambridge University Press.

Buchner, A., Erdfelder, E., Vaterrodt-Plunnecke, B. 1995. Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General* 124:137-60.

Cavanaugh, J.C. 1989. The importance of awareness in memory aging. In (L. Poon, D. Rubin, & B. Wilson, eds) *Everyday Cognition in Adulthood and Late Life*. Cambridge University Press.

Clark, R.E. & Squire, L.R. 1998. Classical conditioning and brain systems: The role of awareness. *Science* 280:77-81.

Cloitre, M. 1997. Conscious and unconscious memory: A model of functional amnesia. In (D. Stein, ed) *Cognitive Science and the Unconscious*. American Psychiatric Press.

Conway, M.A. & Dewhurst, S.A. 1995. The self and recollective experience. *Applied Cognitive Psychology* 9:1-19.

Dalla Barba, G. 2000. Memory, consciousness, and the brain. *Brain & Cognition* 42:20-22.

Duezel, E. 2000. What brain activity tells us about conscious awareness of memory retrieval. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.

Duzel E., Yonelinas A.P., Mangun G.R., Heinze H.J., & Tulving E. 1997. Event related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Sciences of the United States of America* 94:5973-8.

Eich, E. 1984. Memory for unattended events: Remembering with and without awareness. *Memory and Cognition* 12:105-11.

Erdelyi, M. 1984. The recovery of unconscious (inaccessible) memories: Laboratory studies of hypermnnesia. In (G. Bower, ed) *The Psychology of Learning and Motivation*. Academic Press.

Gardiner, J.M. 1988. Functional aspects of recollective experience. *Memory*

and Cognition 16:309-13.

Gardiner, J.M. & Parkin, A.J. 1990. Attention and recollective experience in recognition memory. *Memory and Cognition* 18:579-583.

Gardiner, J.M. 1993. Recognition memory and awareness: An experiential approach. *European Journal of Cognitive Psychology* 5:337-46.

Gardiner, J.M. 1996. On consciousness in relation to memory and learning. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Gardiner, J.M, Ramponi, C. & Richardson-Klavehn, A. 1998. Experiences of remembering, knowing, and guessing. *Consciousness and Cognition* 7:1-26.

Gardiner, J. & Richardson-Klavehn, A. 2000. Remembering and knowing. In (E. Tulving & Craik, F., eds) *The Oxford Handbook of Memory*. Oxford University Press.

Gennaro, R.J. 1992. Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology* 5:333-47.

Gregg, V.H. & Gardiner, J.M. 1994. Recognition memory and awareness: A large effect of study-test modalities on "know" responses following a highly perceptual orienting task. *European Journal of Cognitive Psychology* 6:137-47.

Hamann, S.B. & Squire, L.R. 1997. Intact perceptual memory in the absence of conscious memory. *Behavioral Neuroscience* 111:850-54.

Hirshman, E. & Master, S. 1997. Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory and Cognition* 25:345-351.

Hirst, W. 1989. On consciousness, recall, recognition, and the architecture of memory. In (S. Lewandowsky, J. Dunn, & K. Kirsner, eds) *Implicit Memory: Theoretical Issues*. Lawrence Erlbaum.

Jacoby, L.L. & Witherspoon, D. 1982. Remembering without awareness. *Canadian Journal of Psychology* 36:300-324.

Jacoby, L.L. & Kelley, C.M. 1987. Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin* 13:314-36.

Jacoby, L.L. 1991. A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* 30:513-41.

Jacoby, L.L. & Kelley, C.M. 1991. Unconscious influences of memory: Dissociations and automaticity. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Jacoby, L.L., Toth, J.P. & Yonelinas, A.P. 1993. Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General* 122:139-54.

Jacoby, L.L., Toth, J.P., Yonelinas, A.P. & Debner, J.A. 1994. The relation between conscious and unconscious influences: Independence or redundancy? *Journal of Experimental Psychology: General*.

Jacoby, L.L., Yonelinas, A.P., & Jennings, J.M. 1997. The relation between conscious and unconscious (automatic) influences: A declaration of independence. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Johnson, M.K., Foley, M.A., Suengas, A.G. & Raye, C.L. 1988. Phenomenal characteristics of memories for perceived and imagined autobiographical events.

Journal of Experimental Psychology: General 117:371-76.

Joordens, S. & Merikle, P.M. 1993. Independence or redundancy? Two models of conscious and unconscious influences. Journal of Experimental Psychology: General 122:462-67.

Kelley, C.M. & Jacoby, L.L. 1990. The construction of subjective experience: Memory attributions. Mind and Language 5:49-68.

Kelley, C.M. & Lindsay, D.S. 1996. Conscious and unconscious forms of memory. In (E. Bjork & R. Bjork, eds) *Memory: Handbook of Perception and Cognition*. Academic Press.

Kihlstrom, J.F. 1995. Memory and consciousness: An appreciation of Claparede and "Recognition et Moitte". Consciousness and Cognition 4:379-86.

Kuhlmann, F. 1906. On the analysis of the memory consciousness: A study in the mental imagery and memory of meaningless visual forms. Psychological Review 13:316-48.

Lampinen, J.M., Neuschatz, J.S., & Payne, D.G. 1998. Memory illusions and consciousness: Examining the phenomenology of true and false memories. Current Psychology: Developmental, Learning, Personality, Social 16:181-224.

Larsen, S.F. 1998. What is it like to remember? On phenomenal qualities of memory. In (C. Thompson, J. Read, D. Bruce, D. Payne, & M. Toglia, eds) *Autobiographical and Eyewitness Memory: Theoretical and Applied Perspectives*. Lawrence Erlbaum.

Light, L.L., Singh, A., & Capps, J.L. 1986. Dissociation of memory and awareness in young and older adults. Journal of Clinical & Experimental Neuropsychology 8:62-74.

Lockhart, R.S. 1989. Consciousness and the function of remembered episodes. In (H. Roediger & F. Craik, eds) *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*.

Mandler, G. 1989. Memory: Conscious and unconscious. In (P. Solomon, G. Goethals, C. Kelley, & R. Stephens, eds) *Memory: Interdisciplinary Approaches*. Springer-Verlag.

Moscovitch, M. & Umiltà, C. 1991. Conscious and nonconscious aspects of memory: A neuropsychological framework of modules and central systems. In (R. Lister & H. Weingartner, eds) *Perspectives on Cognitive Neuroscience*. Oxford University Press

Moscovitch, M. 1992. A neuropsychological model of memory and consciousness. In (L. Squire & N. Butters, eds) *Neuropsychology of Memory*. Guilford Press.

Moscovitch, M., Goshen-Gottstein, Y. & Vriezen, E. 1994. Memory without conscious recollection: A tutorial review from a neuropsychological perspective. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance 15*. MIT Press.

Moscovitch, M. 1995. Recovered consciousness: A hypothesis concerning modularity and episodic memory. Journal of Clinical and Experimental Neuropsychology, 17:276-90.

Moscovitch, M. 1995. Models of consciousness and memory. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Moscovitch, M. 2000. Theories of memory and consciousness. In (E. Tulving & F. Craik, eds) *The Oxford Handbook of Memory*. Oxford University Press.

- Natsoulas, T. 1986. Consciousness and memory. *Journal of Mind and Behavior* 7:463-501.
- Oakhill, J. & Kyle, F. 2000. The relation between phonological awareness and working memory. *Journal of Experimental Child Psychology* 75:152-164.
- Paller, K.A., Kutas, M. & McIsaac, H.K. 1995. Monitoring conscious recollection via the electrical activity of the brain. *Psychological Science* 6:107-11.
- Paller, K. 2000. Neural measures of conscious and unconscious memory. *Behavioural Neurology* 12:127-141.
- Penfield, W. 1969. Consciousness, memory, and man's conditioned reflexes. In (H. Hyden, ed) *On the Biology of Learning*. Harcourt, Brace, and World.
- Rajaram, S. & Roediger, H.L. 1997. Remembering and knowing as states of consciousness during retrieval. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Rajaram, S. 1998. The effects of conceptual salience and perceptual distinctiveness on conscious recollection. *Psychonomic Bulletin and Review* 5:71-78.
- Reder, L.M. (ed) 1996. *Implicit Memory and Metacognition*. Lawrence Erlbaum.
- Reingold, E.M. & Toth, P. 1996. Memory dissociations versus task dissociations: A controversy in progress. In (G. Underwood, ed) *Implicit Cognition*. Oxford University Press.
- Richardson-Klavehn, A. & Gardner, J.M. 1996. Cross-modality priming in stem completion reflects conscious memory, but not voluntary memory. *Psychonomic Bulletin and Review* 3:238-44.
- Richardson-Klavehn, A., Gardiner, J.M. & Java, R.I. 1996. Memory: Task dissociations, process dissociations and dissociations of consciousness. In (G. Underwood, ed) *Implicit Cognition*. Oxford University Press.
- Roberts, H.M. 1971. Conscious experiences are a memory process. *Psychological Reports* 29:591-94.
- Roediger, H.L. & Craik, F.I.M. (eds). *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*. Lawrence Erlbaum.
- Ronnberg, J. & Archer, T. 1992. Purposive behaviour in cognition and perception: Considerations of awareness in memory. *Scandinavian Journal of Psychology* 33:86-91.
- Rugg, M.D. 1995. Memory and consciousness: A selective review of issues and data. *Neuropsychologia* 33:1131-1141.
- Schacter, D.L. 1987. Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:501-18.
- Schacter, D.L. 1989. On the relation between memory and consciousness: Dissociable interactions and conscious experience. In (H. Roediger & F. Craik, eds) *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*.
- Schacter, D.L., Bowers, J. & Booker, J. 1989. Intention, awareness, and implicit memory: The retrieval intentionality criterion. In (S. Lewandowsky, J. Dunn, & K. Kirsner, eds) *Implicit Memory: Theoretical Issues*. Lawrence Erlbaum.

Schacter, D.L. 1992. Consciousness and awareness in memory and amnesia: Critical issues. In (A. Milner & M. Rugg, eds) *The Neuropsychology of Consciousness*. Academic Press.

Schacter, D.L. 1995. Implicit memory: A new frontier for cognitive neuroscience. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Schacter, D.L. 1998. Memory and awareness. *Science* 280:59-60.

Stolz, J. & Merikle, P. 2000. Conscious and unconscious influences of memory: Temporal dynamics. *Memory* 8:333-343.

Toth, J.P., Lindsay, D.S, & Jacoby, L.L. 1992. Awareness, automaticity, and memory dissociations. In (L. Squire & N. Butters, eds) *Neuropsychology of Memory*. Guilford Press.

Toth, J.P. & Reingold, E.M. 1996. Beyond perception: Conceptual contributions to unconscious influences of memory. In (G. Underwood, ed) *Implicit Cognition*. Oxford University Press.

Tulving, E. 1985. Memory and consciousness. *Canadian Psychology* 26:1-12.

Tulving, E. 1987. Multiple memory systems and consciousness. *Human Neurobiology* 6:67-80.

Tulving, E. 1993. Varieties of consciousness and levels of awareness in memory. In (A. Baddeley & L. Weiskrantz, eds) *Attention: Selection, Awareness, and Control*. Oxford University Press.

Underwood, G. 1979. Memory systems and conscious processes. In (G. Underwood & R. Stevens, eds) *Aspects of Consciousness, Volume 1*. Academic Press.

Verfaellie, M. & Keane, M.M. 1997. The neural basis of aware and unaware forms of memory. *Seminars in Neurology* 17:153-61.

Wheeler, M.A. Stuss, D.T. & Tulving, E. 1997. Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin* 121:331-54.

Wippich, W. 1992. Implicit and explicit memory without awareness. *Psychological Research* 54:212-24.

6.2g Consciousness and Learning

Berry, D.C. 1994. Implicit learning: Twenty-five years on. A tutorial. In (C. Umiltà and M. Moscovitch, eds) *Consciousness and Unconscious Information Processing: Attention and Performance 15*. MIT Press.

Berry, D.C. 1997. *How Implicit is Implicit Learning?*. Oxford University Press.

Berry, D.C. & Dienes, Z. (eds) 1993. *Implicit Learning: Theoretical and Empirical Issues*. Lawrence Erlbaum Associates.

Brody, N. 1989. Unconscious learning of rules: Comment on Reber's analysis of implicit learning. *Journal of Experimental Psychology: General* 118:236-238.

Carlson, R.A. & Dulany, D.E. 1985. Conscious attention and abstraction in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11:45-58.

Cleeremans, A. 1993. Mechanisms of implicit learning: Connectionist models of sequence processing.

Conway, M.A., Collins, A.F., Anderson, S.J., & Cohen, G. 1998. Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General*.

Dienes, Z. & Berry, D. 1997. Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review* 4:3-23.

Dulany, D.E. 1968. Awareness, rules, and propositional control: A confrontation with S-R behavior theory. In (T. Dixon & D. Horton, eds) *_Verbal Behavior and General Behavior Theory_*. Prentice-Hall.

Dulany, D.E., Carlson, R.A. & Dewey, G.I. 1984. A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General* 113:541-555.

Dulany, D.E., Carlson, R.A., Dewey, G.I. 1985. On consciousness in syntactic learning and judgment: A reply to Reber, Allen, and Regan. *Journal of Experimental Psychology: General* 114:25-32.

Furedy, J., Damke, B., & Boucsein, W. 2000. Revisiting the learning-without-awareness question in human Pavlovian autonomic conditioning: Focus on extinction in a dichotic listening paradigm. *Integrative Physiological & Behavioral Science* 35:17-34.

Gardiner, J.M. 1996. On consciousness in relation to memory and learning. In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.

Kihlstrom, J.F. 1996. Perception without awareness of what is perceived, learning without awareness of what is learned. In (M. Velmans, ed) *_The Science of Consciousness_*. Routledge.

Manns, J., Clark, R., & Squire, L. 2000. Awareness predicts the magnitude of single-cue trace eyeblink conditioning. *Hippocampus* 10:181-186.

Marton, F. & Booth, S.A. 1997. *_Learning and Awareness_*. Lawrence Erlbaum.

Neal, A. & Hesketh, B. 1997. Episodic knowledge and implicit learning. *Psychonomic Bulletin and Review* 4:24-37.

Neal, A. & Hesketh, B. 1997. Future directions for implicit learning: Toward a clarification of issues associated with knowledge representation and consciousness. *Psychonomic Bulletin and Review* 4:73-78.

Perruchet, P., Gallego, J. & Savy, I. 1990. A critical reappraisal of the evidence for unconscious abstraction of deterministic rules in complex experimental situations. *Cognitive Psychology* 22:493-516.

Perruchet, P. & Pacteau, C. 1990. Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge. *Journal of Experimental Psychology: General* 119:264-75.

Perruchet, P., Vinter, A., & Gallego, J. 1997. Implicit learning shapes new conscious percepts and representations. *Psychonomic Bulletin & Review* 4:43-48.

Reber, A.S. 1967. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* 6:855-863.

Reber, A.S. 1989. Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* 118:219-35.

Reber, A.S. 1993. *_Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious_*. Oxford University Press.

Reber, A.S. 1997. How to differentiate implicit and explicit modes of acquisition. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Reber, A.S., Allen, R., & Regan, S. 1985. Syntactical learning and judgment, still unconscious and still abstract: Comment on Dulany, Carlson, and Dewey. *Journal of Experimental Psychology: General* 114:17-24.

Shanks, D.R., Green, R.E.A., & Kolodny, J.A. 1994. A critical examination of the evidence for unconscious (implicit) learning. In (C. Umiltà and M. Moscovitch, eds) *_Consciousness and Unconscious Information Processing: Attention and Performance 15_*. MIT Press.

Shanks, D.R. & St. John, M.F. 1994. Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* 17:367-447.

Sno, H. 2000. Déjà vu and jamais vu. In (G. Berrios & J. Hodges, eds) *_Memory Disorders in Psychiatric Practice_*. Cambridge University Press.

Stadler, M.A. & Roediger, H.L. 1998. The question of awareness in research on implicit learning. In (M. Stadler & P. Frensch, eds) *_Handbook of Implicit Learning_*. Sage Publications

Stadler, M.A., & Frensch, P.A. 1998. *_Handbook of Implicit Learning_*. Sage Publications.

Whittlesea, B.W.A. & Dorken, M.D. 1997. Implicit learning: Indirect, not unconscious. *Psychonomic Bulletin & Review* 4:63-67.

Wong, P.S., Bernat, E., Bunce, S., & Shevrin, H. 1997. Brain indices of nonconscious associative learning. *Consciousness and Cognition* 6:519-544.

6.2h Consciousness and Metacognition

Brown, R. & McNeill, D. 1966. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior* 5:325-37.

Brown, S. 2000. Tip-of-the-tongue phenomena: An introductory phenomenological analysis. *Consciousness & Cognition* 9:516-537.

Hart, J.T. 1965. Memory and the feeling-of-knowing experience. *Journal of Educational Psychology* 56:208-16.

Johnson, M.K. 1988. Reality monitoring: An experimental phenomenological approach. *Journal of Experimental Psychology: General* 117:390-94.

Johnson, M.K. 1991. Reflection, reality monitoring, and the self. In (R. Kunzendorf, ed) *_Mental Imagery_*. Plenum Press.

Johnson, M.K. & Reeder, J.A. 1997. Consciousness as meta-processing. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Kahan, T.L. & LaBerge, S. 1994. Lucid dreaming as metacognition: Implications for cognitive science. *Consciousness & Cognition* 3:246-64.

Mangan, B. 2000. What feeling is the "feeling of knowing?" *Consciousness & Cognition* 9:538-544.

Metcalfe, J. & Shimamura, P. 1994. *_Metacognition: Knowing about Knowing_*. MIT Press.

Nelson, T.O. 1996. Consciousness and metacognition. *American Psychologist*

51:102-16.

Nelson, T.O. 1992. Metacognition: Core Readings. Allyn and Bacon.

Nisbett, R. & Wilson, T. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review* 84:231-59.

Otani, H. & Hodge, M. 1991. mechanisms of feelings of knowing: The role of elaloration and familiarity. *Psychological Record* 41:523-35.

Reder, L.M. (ed) 1996. Implicit Memory and Metacognition. Lawrence Erlbaum.

Reder, L.M. & Schunn, C.D. 1996. Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In (L. Reder, ed) Implicit Memory and Metacognition. Lawrence Erlbaum.

Ricciardelli, L.A. 1993. Two components of metalinguistic awareness: Control of linguistic processing and analysis of linguistic knowledge. *Applied Psycholinguistics* 14:349-367.

Rosenthal, D.M. 1998. Consciousness and metacognition. In (D. Sperber, ed) Metarepresentation. Oxford University Press.

Wegner, D.M. 1989. White Bears and Other Unwanted Thoughts: Suppression, Obsession, and the Psychology of Mental Control. Penguin.

Wegner, D.M. 1997. Why the mind wanders. In (J. Cohen & J. Schooler, eds) Scientific Approaches to Consciousness. Lawrence Erlbaum.

White, P. 1980. Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review* 87:105-12.

White, P. 1983. Knowing our own minds: Conscious awareness and verbal reports. In (G. Underwood, ed) Aspects of Consciousness, Volume 3: Awareness and Self-Awareness. Academic Press.

Wilson, T.D. 1997. The psychology of metapsychology. In (J. Cohen & J. Schooler, eds) Scientific Approaches to Consciousness. Lawrence Erlbaum.

6.2i Consciousness and Control

Baars, B.J. 1987. What is conscious in the control of action? A modern ideomotor theory of voluntary action. In (D. Gorfein & R. Hoffman, eds) Learning and Memory: The Ebbinghaus Centennial Symposium. Lawrence Erlbaum.

Baars, B.J. 1992. Experimental Slips and Human Error: Exploring the Architecture of Volition. Plenum Press.

Baars, B.J. 1993. Why volition is a foundation issue for psychology. *Consciousness and Cognition* 2:281-309.

Bargh, J.A. 1996. Automaticity in action: The unconscious as repository of chronic goals and motives. In (P. Gollwitzer & J. Bargh, eds) The Psychology of Action: Linking Cognition and Motivation to Behavior. Guilford.

Bargh, J.A. 1996. Automaticity in social psychology. In (E. Higgins & A. Kruglanski, eds) Social Psychology: Handbook of Basic Principles. Guilford.

Bargh, J.A. 1994. The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In (R. Wyer & T. Srull, eds) Handbook of Social Cognition. Lawrence Erlbaum.

Bayles, G.H. & Cleary, P.J. 1986. The role of awareness in the control of

frontalis muscle activity. *Biological Psychology* 22:23-35.

Carr, T.H., McCauley, C., Sperber, R.D., & Parmelee, C.M. 1982. Words, pictures, and priming: On semantic activation, conscious identification, and the automaticity of information processing. *Journal of Experimental Psychology: Human Perception & Performance* 8:757-777.

Daprati, E., Franck, N., Georgieff, N.,; Proust, J. 1997. Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition* 65:71-86.

Delabarre, E.B. 1911. Volition and motor consciousness: Theory. *Psychological Bulletin* 8:378-82.

Delabarre, E.B. 1912. Volition and motor consciousness: Theory. *Psychological Bulletin* 9:409-13.

Delabarre, E.B. 1913. Volition and motor consciousness. *Psychological Bulletin* 10:441-44.

Dewan, E.M. 1976. Consciousness as an emergent causal agent in the context of control system theory. In (G. Globus, G. Maxwell, & I. Savodnik, eds) *_Consciousness and the Brain_*. Plenum Press.

Gott, P.S., Hughes, E.C. & Whipple, K. 1984. Voluntary control of two lateralized conscious states: Validation of electrical and behavioral studies. *Neuropsychologia* 22:65-72.

Gordon, A.M. & Rosenbaum, D.A. 1984. Conscious and subconscious arm movements: Application of signal detection theory to motor control. *Bulletin of the Psychonomic Society* 22:214-216.

Gray, J.A. 1998. Abnormal contents of consciousness: The transition from automatic to controlled processing. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Hilgard, E.R. 1979. Consciousness and control: Lessons from hypnosis. *Australian Journal of Clinical & Experimental Hypnosis* 7:103-15.

Horowitz, M.J. & Stinson, C.H. 1995. Consciousness and processes of control. *Journal of Psychotherapy Practice and Research* 4:123-139.

Jacoby, L.L. 1991. A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language* 30:513-41.

Jacoby, L.L., Ste-Marie, D. & Toth, J.P. 1993. Redefining automaticity: Unconscious influences, awareness, and control. In (A. Baddeley & L. Weiskrantz, eds) *_Attention: Selection, Awareness, and Control_*. Oxford University Press.

Kamiya, J. 1968. Conscious control of brain waves. *Psychology Today* 1:56-60.

Kimble, G.A. & Perlmutter, L.C. 1970. The problem of volition. *Psychological Review* 77:361-84.

Langer, E.J. 1992. Matters of mind: Mindfulness/mindlessness in perspective. *Consciousness and Cognition* 1:289-305.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8:529-566.

Oswald, M. & Gadenne, V. 2000. Are controlled processes conscious? In (W. Perrig & A. Grob, eds) *_Control of Human Behavior, Mental Processes, and*

Consciousness: Essays in Honor of the 60th Birthday of August Flammer_.
Lawrence Erlbaum Associates.

Plotkin, W.B. 1976. On the self-regulation of the occipital alpha rhythm:
Control strategies, states of consciousness, and the role of physiological
feedback. *Journal of Experimental Psychology: General* 105:66-99.

Plotkin, William B. 1981. A rapprochement of the operant-conditioning and
awareness views of biofeedback training: The role of discrimination in
voluntary control. *Journal of Experimental Psychology: General* 110:415-428.

Posner, M. & Snyder, C.R.R. 1975. Attention and cognitive control. In
(R. Solso, ed) *Information Processing and Cognition: The Loyola Symposium*_.
Lawrence Erlbaum.

Raichle, M.E. 1997. Automaticity: From reflective to reflexive information
processing. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) *Cognition,
Computation, and Consciousness*_. Oxford University Press.

Schneider, W.E. & Shiffrin, R.M. 1977. Controlled and automatic human
information processing: I. Detection, search, and attention. *Psychological
Review* 84:1-66.

Shiffrin, R.M. & Schneider, W.E. 1977. Controlled and automatic human
information processing: II. Perceptual learning, automatic attending, and a
general theory. *Psychological Review* 84:128-90.

Tzelgov, J. 1997. Specifying the relations between automaticity and
consciousness: A theoretical note. *Consciousness and Cognition* 6:441-51.

Tzelgov, J., Porat, Z. & Henik, A. 1997. Automaticity and consciousness: Is
perceiving the word necessary for reading it? *American Journal of
Psychology* 110:429-48.

Tzelgov, J. 1997. Automatic but conscious: That is how we act most of the
time. In (R. Wyer, ed) *The Automaticity of Everyday Life*_. Lawrence Erlbaum.

Uleman, J.S. 1987. Consciousness and control: The case of spontaneous trait
inferences. *Personality and Social Psychology Bulletin* 13:337-54.

Umiltà, C. 1988. The control operations of consciousness. In (A. Marcel & E.
Bisiach, eds) *Consciousness in Contemporary Science*_. Oxford University Press.

White, W.A. 1920. Extending the field of conscious control. *Mental Hygiene*
4:857-66.

Zelazo, P.D. & Frye, D. 1997. Cognitive complexity and control: A theory of
the development of deliberate reasoning and intentional action. In
(M. Stamenov, ed) *Language Structure, Discourse, and the Access to
Consciousness*_. John Benjamins.

6.2j Consciousness and Imagery

Ahsen, A. 1991. Imagery and consciousness: Putting together poetic, mythic and
social realities. *Journal of Mental Imagery* 15:63-97.

Ahsen, A. 1991. A second report on AA-VVIQ: Role of vivid and unvivid images
in consciousness research. *Journal of Mental Imagery* 15:1-31.

Ahsen, A. 1993. Imagery paradigm: Imaginative consciousness in the
experimental and clinical setting. *Journal of Mental Imagery*.

Arnheim, R. 1994. Consciousness: An island of images. *Journal of Theoretical and Philosophical Psychology* 14:121-27.

Bichowsky, F.R. 1926. The mechanism of consciousness: Images. *American Journal of Psychology* 37:557-564.

Frick, R.W. 1987. A dissociation of conscious visual imagery and visual short-term memory. *Neuropsychologia* 25:707-12.

Hampson, P.J. & Morris, P.E. 1990. Imagery, consciousness, and cognitive control: The BOSS model reviewed. In (P. Hampson, D. Marks, & J. Richardson, Eds) *_Imagery: Current Developments_*. Routledge.

Hebb, D.O. 1968. Concerning imagery. *Psychological Review* 75:466-77.

Horne, P.V. 1993. The nature of imagery. *Consciousness and Cognition* 2:58-82.

Hubbard, T.L. 1996. The importance of a consideration of qualia to imagery and cognition. *Consciousness and Cognition* 3:327-58.

Ishai, A. & Sagi, D. 1998. Visual imagery and visual perception: The role of memory and conscious awareness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.

Krellenstein, M.F. 1995. Unsolvable problems, visual imagery, and explanatory satisfaction. *Journal of Mind and Behavior* 16:235-54.

Kunzendorf, R.G. 1990. The causal efficacy of consciousness in general, imagery in particular: A materialist perspective. In (R. Kunzendorf, ed) *_Mental Imagery_*. Plenum Press.

Kunzendorf, R.G., Justice, M., & Capone, D. 1997. Conscious images as "centrally excited sensations": A developmental study of imaginal influences on the ERG. *Journal of Mental Imagery* 21:155-66.

Lehmann, D., Henggler, B., Koukkan, M. & Michel, M. 1993. Source localization of brain electric field frequency bands during conscious, spontaneous visual imagery and abstract thought. *Cognitive Brain Research* 1:203-20.

Mandler, G. 1984. Consciousness, imagery, and emotion -- with special reference to autonomic imagery. *Journal of Mental Imagery* 8:87-94.

Marks, D.F. 1977. Imagery and consciousness: A theoretical review from an individual differences perspective. *Journal of Mental Imagery* 1:275-90.

Marks, D.F. 1983. Imagery and consciousness: A theoretical review. In (A. Sheikh, ed) *_Imagery: Current Theory, Research, and Application_*. Wiley.

Marks, D.F. 1990. On the relationship between imagery, body, and mind. In (P. Hampson, D. Marks, & J. Richardson, eds) *_Imagery: Current Developments_*. Routledge.

Mavromatis, A. 1987. On shared states of consciousness and objective imagery. *Journal of Mental Imagery* 11:125-30.

Morris, P.E. & Hampson, P.J. 1983. *_Imagery and Consciousness_*. Academic Press.

Newton, N. 1982. Experience and imagery. *Southern Journal of Philosophy* 21:475-87.

Richardson, A. 2000. Individual differences in visual imagination imagery. In (R. Kunzendorf & B. Wallace, eds) *_Individual Differences in Conscious Experience_*. John Benjamins.

Sheehan, P.W. & Lewis, S.E. 1974. Subjects' reports of confusion in consciousness and the arousal of imagery. *Perceptual and Motor Skills* 38:731-34.

6.2k Consciousness and Emotion

Cioffi, D. 1991. Sensory awareness versus sensory impression: Affect and attention interact to produce somatic meaning. *Cognition and Emotion* 5:275-94.

Damasio, A. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.

DeLancey, C. 1996. Emotion and the function of consciousness. *Journal of Consciousness Studies* 3:492-99.

Dimberg, U., Thunberg, M., & Elmehed, K. 2000. Unconscious facial reactions to emotional facial expressions. *Psychological Science* 11:86-89.

Ellis, R.D. 1995. *Questioning Consciousness: The Interplay of Imagery, Cognition, and Emotion in the Human Brain*. John Benjamins.

Ellis, R. 2000. *The Caldron of Consciousness: Motivation, Affect and Self-organization*. John Benjamins.

Forgas, J. & Ciarrochi, J. 2000. Affect infusion and affect control: The interactive role of conscious and unconscious processing strategies in mood management. In (Y. Rossetti & A. Revonsuo, eds) *Beyond Dissociation: Interaction between Dissociated Implicit and Explicit Processing*. John Benjamins.

Gray, J.A. 1999. Cognition, emotion, conscious experience and the brain. In (T. Dalgleish & M.J. Powers, eds) *Handbook of Cognition and Emotion*. Wiley.

Heilman, K. 2000. Emotional experience: A neurological model. In (R. Lane, L. Nadel, & G. Ahern, J. Allen, & A. Kaszniak, eds) *Cognitive Neuroscience of Emotion*. Oxford University Press.

Katz, J.M. 1983. Altered states of consciousness and emotion. *Imagination, Cognition and Personality* 2:37-50.

Kunst-Wilson, W.R. & Zajonc, R.B. 1980. Affective discrimination of stimuli that cannot be recognized. *Science* 207:557-58.

Lane, R.D. Reiman, E., Ahern, G., Schwartz, G.E., & Yun, L. 1998. Anterior cingulate cortex participates in the conscious experience of emotion. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Lane, R. 2000. Neural correlates of conscious emotional experience. In (R. Lane, L. Nadel, & G. Ahern, J. Allen, & A. Kaszniak, eds) *Cognitive Neuroscience of Emotion*. Oxford University Press.

Lane, R.D., Ahern, G.L., Schwartz, G.E. & Kaszniak, A.W. 1997. Is alexithymia the emotional equivalent of blindsight? *Biological Psychiatry* 42:834-44.

Ledoux, J.E. 1995. In search of an emotional system in the brain: Leaping from fear to emotion and consciousness. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Lewis, M. 1997. The self in self-conscious emotions. In (J. Snodgrass, R. Thompson, eds) *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.

Lishman, W.A. 1971. Emotion, consciousness, and will after brain bisection in man. *Cortex* 7:181-92.

Matsumoto, D. & Lee, M. 1993. Consciousness, volition, and the neuropsychology of facial expressions of emotion. *Consciousness and Cognition* 2:237-54.

Morris, J.S., Ohman, A., & Dolan, R.J. 1998. Conscious and unconscious emotional learning in the human amygdala. *Nature* 393:467-470.

Niedenthal, P.M. 1990. Implicit perception of affective information. *Journal of Experimental Social Psychology* 26:505-27.

Ochsner, K. 2000. Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology: General* 129:242-261.

Ohman, A., Flykt, A., & Lundqvist, D. 2000. Unconscious emotion: Evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. In (R. Lane, L. Nadel, & G. Ahern, eds) *_Cognitive Neuroscience of Emotion_*. Oxford University Press.

Peper, M. 2000. Awareness of emotions: A neuropsychological perspective. In (R. Ellis, ed) *_The Caldron of Consciousness: Motivation, Affect and Self-organization_*. John Benjamins Publishing Company.

Pratto, F. 1994. Consciousness and automatic evaluation. In (P. Niedenthal & S. Kitayama, eds) *_The Heart's Eye: Emotional Influences in Perception and Attention_*. Academic Press.

Reiman, E.M., Lane, R.D., Ahern, G.L., Schwartz, G.E. 1996. Positron emission tomography, emotion, and consciousness. In (S. Hamreoff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Rolls, E.T. 1995. A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In (M. Gazzaniga, eds) *_The Cognitive Neurosciences_*. MIT Press.

Safran, J.D. & Greenberg, L.S. 1987. Affect and the unconscious: A cognitive perspective. In (R. Stern, ed) *_Theories of the Unconscious and Theories of the Self_*. Analytic Press.

Wakefield, J.C. 1991. Why emotions can't be unconscious: An exploration of Freud's essentialism. *Psychoanalysis and Contemporary Thought* 14:29-67.

Weiskrantz, L. 2000. Blindsight: Implications for the conscious experience of emotion. In (R. Lane, L. Nadel, & G. Ahern, eds) *_Cognitive Neuroscience of Emotion_*. Oxford University Press.

Zajonc, R. 2000. Feeling and thinking: Closing the debate over the independence of affect. In (J. Forgas, ed) *_Feeling and Thinking: The Role of Affect in Social Cognition_*. Cambridge University Press.

6.21 Consciousness, Sleep, and Dreaming -----

Arden, J.B. 1996. *_Consciousness, Dreams, and Self: A Transdisciplinary Approach_*. Psychosocial Press.

Bentley, E. 2000. *_Awareness: Biorhythms, Sleep and Dreaming_*. Routledge.

Bosinelli, M. 1995. Mind and consciousness during sleep. *Behavioural Brain Research* 69:195-201.

Broughton, R.J. 1982. Human consciousness and sleep/waking rhythms: A review and some neuropsychological considerations. *Journal of Clinical Neuropsychology* 4:193-218.

Combs, A. & Krippner, S. 1998. Dream sleep and waking reality: A dynamical view. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Flanagan, O. 1997. Prospects for a unified theory of consciousness or, what dreams are made of. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Foulkes, D. 1990. Dreaming and consciousness. *European Journal of Cognitive Psychology* 2:39-55.

Gackenbach, J. & LaBerge, S. 1988. *Conscious Mind, Sleeping Brain: Perspectives on Lucid Dreaming*. Plenum Press.

Green, C. & McGreery, C. 1994. *Lucid Dreaming: The Paradox of Consciousness During Sleep*. Routledge.

Hearne, K.M. 1992. ProLucid dreaming, lucid dreams, and consciousness. *Journal of Mental Imagery* 16:119-123.

Hobson, J.A. & Strickgold, R. 1995. The conscious state paradigm: A neurocognitive approach to waking, sleeping, and dreaming. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Hobson, J.A. 1998. The conscious state paradigm: A neuropsychological analysis of waking, sleeping, and dreaming. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Hobson, J., Pace-Schott, E., & Stickgold, R. 2000. Consciousness: Its vicissitudes in waking and sleep. In (M. Gazzaniga, ed) *The New Cognitive Neurosciences: 2nd Edition*. MIT Press.

Jones, B.E. 1998. The neural basis of consciousness across the sleep-waking cycle. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *Consciousness: At the Frontiers of Neuroscience*. Lippincott-Raven.

Kahan, T.L. & LaBerge, S. 1994. Lucid dreaming as metacognition: Implications for cognitive science. *Consciousness & Cognition* 3:246-264.

Kahan, T.L. & LaBerge, S. 1996. Cognition and metacognition in dreaming and waking: Comparisons of first and third-person ratings. *Dreaming* 6:235-249.

Kahn, D., Pace-Schott, E.F. & Hobson, J.A. 1997. Consciousness in waking and dreaming: The roles of neuronal oscillation and neuromodulation in determining similarities and differences. *Neuroscience* 78:13-38.

Khan, D., Krippner, S., & Combs, A. 2000. Dreaming and the self-organizing brain. *Journal of Consciousness Studies* 7:4-11.

King, C.D. 1947. Dream and the problem of consciousness. *Journal of General Psychology* 37:15-24.

Kleitman, N. 1957. Sleep, wakefulness, and consciousness. *Psychological Bulletin* 54:354-359.

LaBerge, S. 1985. *Lucid Dreaming*. J.P. Tarcher.

LaBerge, S., Levitan, L., & Dement, W.C. 1986. Lucid dreaming: Physiological correlates of consciousness during REM sleep. *Journal of Mind and Behavior*

7:251-258.

LaBerge, S. 1990. Lucid dreaming: Psychophysiological studies of consciousness during REM sleep. In (R. Bootsen, J. Kihlstrom, & D. Schacter, eds) *Sleep and Cognition*. American Psychological Association Press.

LaBerge, S. 1998. Dreaming and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

LaBerge, S. & DeGracia, D. 2000. Varieties of lucid dreaming experience. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Lindsley, D.B. 1960. Attention, consciousness, sleep, and wakefulness. In (H. Magoun & V. Hall, eds) *Handbook of Physiology. Section I: Neurophysiology*. American Physiological Society.

Makeig, S., Jung, T., & Sejnowski, T. 2000. Awareness during drowsiness: Dynamics and electrophysiological correlates. *Canadian Journal of Experimental Psychology* 54:266-273.

Monnier, M. 1952. Experimental work on sleep and other variations of consciousness. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Third Conference*. Josiah Macy Foundation.

Moorcroft, W. & Breitenstein, J. 2000. Awareness of time during sleep. *Annals of Medicine* 32:236-238.

Munglani, R. & Jones, J.G. 1992. Sleep and general anaesthesia as altered states of consciousness. *Journal of Psychopharmacology* 6:399-409.

Pare, D. & Llinas, R. 1995. Conscious and pre-conscious processes as seen from the standpoint of sleep-waking cycle neurophysiology. *Neuropsychologia* 33:1155-1168.

Revonsuo, A. 1995. Consciousness, dreams and virtual realities. *Philosophical Psychology* 8:35-58.

Simon, C.W. & Emmons, W. 1956. EEG, consciousness, and sleep. *Science* 124:1066-1069.

Stoyva, J. & Kamiya, J. 1968. Electrophysiological studies of dreaming as the prototype of a new strategy in the study of consciousness. *Psychological Review* 75:192-205.

6.2m Consciousness in Psychoanalysis

Balint, E. 1987. Memory and consciousness. *International Journal of Psychoanalysis* 68:475-483.

Barr, H. & Langs, R. 1972. The psychoanalytic theory of consciousness. In *LSD: Personality and experience*. Wiley-Interscience.

Bouveresse, J. 1995. *Wittgenstein Reads Freud: The Myth of the Unconscious*. Princeton University Press.

Bower, G.H. 1990. Awareness, the unconscious, and repression: An experimental psychologist's perspective. In (J. Singer, ed) *Repression and Dissociation*. University of Chicago Press.

Brakel L.W. 1989. Negative hallucinations, other irretrievable experiences and two functions of consciousness. *International Journal of Psychoanalysis* 70:461-89.

- Burston, D. 1986. The cognitive and dynamic unconscious: A critical and historical perspective. *Contemporary Psychoanalysis* 22:133-57.
- Chang, S.C. 1978. The psychology of consciousness. *American Journal of Psychotherapy* 32:105-116.
- Curtis, R. 1992. A process view of consciousness and the "self": Integrating a sense of connectedness with a sense of agency. *Psychological Inquiry*, 3:29-32.
- Eagle, M.N. 1987. The psychoanalytic and the cognitive unconscious. In (R. Stern, ed) *Theories of the Unconscious and Theories of the Self*. Analytic Press.
- Epstein, S. 1994. Integration of the cognitive and the psychodynamic unconscious. *American Psychologist* 49:409-24.
- Erdelyi, M.H. 1988. Issues in the study of unconscious and defense processes. In (M. Horowitz, ed) *Psychodynamics and Cognition*. University of Chicago Press.
- Foulkes, D. 1964. Theories of dream formation and recent studies of sleep consciousness. *Psychological Bulletin* 62:236-47.
- Globus, G.G. 1974. The problem of consciousness. *Psychoanalysis and Contemporary Science* 3:40-69.
- Haldane, J. 1988. Psychoanalysis, cognitive psychology and self-consciousness. In (P. Clark & C. Wright, eds) *Mind, Psychoanalysis and Science*. Blackwell.
- Herzog, P.S. 1991. *Conscious and Unconscious: Freud's Dynamic Distinction Reconsidered*. International Universities Press.
- Horowitz, M.J. & Stinson, C.H. 1995. Consciousness and processes of control. *Journal of Psychotherapy Practice and Research* 4:123-139.
- Joseph E.D. 1987. The consciousness of being conscious. *Journal of the American Psychoanalytic Association* 35:5-22.
- Klein, G. 1959. Consciousness in psychoanalytic theory. *Journal of the American Psychoanalytic Association* 7:5-34.
- Kubie, L.S. 1954. Psychiatric and psychoanalytic considerations of the problem of consciousness. In (J. Delafresnaye, ed) *Brain Mechanisms and Consciousness*. Blackwell.
- Levy, D. 1996. *Freud among the Philosophers: The Psychoanalytic Unconscious and its Philosophical Critics*. Yale University Press.
- Masek, R. 1989. The overlooked problem of consciousness in psychoanalysis: Pierre Janet revisited. *Humanistic Psychologist* 17:274-279.
- MacIntyre, A.C. 1958. *The Unconscious: A Conceptual Study*. London.
- Miller, L. 1997. Freud and consciousness: The first one hundred years of neuropsychodynamics in theory and clinical practice. *Seminars in Neurology* 17:171-77.
- Moraglia, G. 1991. The unconscious in information processing and analytical psychology. *Journal of Analytical Psychology* 36:27-36.
- Natsoulas, T. 1992. Toward an improved understanding of Sigmund Freud's conception of consciousness. *Journal of Mind & Behavior* 13:171-92.

Natsoulas, T. 1984-1996. Freud and consciousness I-XI. *Psychoanalysis and Contemporary Thought* 7:195-232, 8:183-220, 12:97-123, 12:619-62, 14:69-108, 15:305-48, 16:67-101, 16:597-631, 19:461-94.

Natsoulas, T. 1995. A rediscovery of Sigmund Freud. *Consciousness and Cognition* 4:300-322.

Olds, D.D. 1992. Consciousness: A brain-centered, informational approach. *Psychoanalytic Inquiry* 12:419-44.

Opatow, B. 1997. The real unconscious: Psychoanalysis as a theory of consciousness. *Journal of the American Psychoanalytic Association* 45:865-90.

Orbach, I. 1995. *The Hidden Mind: Psychology, Psychotherapy, and Unconscious Processes*. Wiley.

Piaget, J. 1973. The affective unconscious and the cognitive unconscious. *Journal of the American Psychoanalytic Association* 21:249-261.

Power, M.J. & Brewin, C.R. 1991. From Freud to cognitive science: A contemporary account of the unconscious. *British Journal of Clinical Psychology* 30:289-310.

Power, M.J. 1997. Conscious and unconscious representations of meaning. In (M. Power & C. Brewin, eds) *The Transformation of Meaning in Psychological Therapies: Integrating Theory and Practice*. John Wiley.

Rapaport, D. 1951. Consciousness: A Psychopathological and psychodynamic view. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Second Conference*. Josiah Macy Foundation.

Rosenblatt A.D. & Thickstun J.T. 1994. Intuition and consciousness. *Psychoanalytic Quarterly* 63:696-714.

Rubinfine, D.L. 1973. Notes toward a theory of consciousness. *International Journal of Psychoanalytic Psychotherapy* 2:391-410.

Ryle, A. 1994. Consciousness and psychotherapy. *British Journal of Medical Psychology* 67:115-23.

Schimek, J.G. 1975. A critical re-examination of Freud's concept of unconscious mental representation. *International Review of Psychoanalysis* 2:171-87.

Shevrin, H. 1990. Subliminal perception and repression. In (J. Singer, ed) *Repression and Dissociation: Implications for Personality Theory, Psychopathology, and Health*. University of Chicago Press.

Shevrin, H. 1992. The Freudian unconscious and the cognitive unconscious: Identical or fraternal twins? In (J. Barron, M. Eagle, & D. Wolitzky, eds) *Interface of Psychoanalysis and Psychology*. American Psychological Association.

Shevrin, H., Williams, W.J., Marshall, R.E., & Brakel, L.A. 1992. Event-related potential indicators of the dynamic unconscious. *Consciousness and Cognition* 1:340-66.

Shevrin, H., Bond, J., Brakel, L., Hertel, R. & Williams, W. 1996. *Conscious and Unconscious Processes: Psychodynamic, Cognitive, and Neurophysiological Convergences*. Guilford Press.

Shevrin, H. 1998. The Freud-Rapaport theory of consciousness. In (R. Bornstein & J. Masling, eds) *Empirical Perspectives on the Psychoanalytic*

Unconscious_. American Psychological Association.

Slipp, S. 2000. Subliminal stimulation research and its implications for psychoanalytic theory and treatment. *Journal of the American Academy of Psychoanalysis* 28:305-320.

Smith, D. 2000. Freudian science of consciousness: Then and now. *Neuro-psychoanalysis* 2:38-45.

Solomon, R.C. 1974. Freud and "unconscious motivation". *Journal for the Theory of Social Behaviour* 4:191-216.

Solms, M. 1997. What is consciousness? *Journal of the American Psychoanalytic Association* 45:681-703.

Spence, D.P. & Holland, B. 1962. The restricting effects of awareness: A paradox and an explanation. *Journal of Abnormal and Social Psychology* 64:163-74.

Stein, D.J. (ed) 1997. *_Cognitive Science and the Unconscious_*. American Psychiatric Press.

Strauss, A. 1955. Unconscious mental processes and the psychosomatic concept. *International Journal of Psychoanalysis* 36:307-19.

van der Waals, E.G. 1949. The psycho-analytical and the phenomenological concept of consciousness. *International Journal of Psychoanalysis* 30:207.

Wakefield, J.C. 1990. Why instinctual impulses can't be unconscious: An exploration of Freud's cognitivism. *Psychoanalysis and Contemporary Thought* 13:265-88.

Wakefield, J.C. 1991. Why emotions can't be unconscious: An exploration of Freud's essentialism. *Psychoanalysis and Contemporary Thought* 14:29-67.

Weinberger, J.& Weiss, J. 1997. Psychoanalytic and cognitive conceptions of the unconscious. In (D. Stein, ed) *_Cognitive Science and the Unconscious_*. American Psychiatric Press.

Westen, D. 1992. The cognitive self and the psychoanalytic self: Can we put our selves together? *Psychological Inquiry* 3:1-13.

Woody, J.M. & Phillips, J. 1995. Freud's project for a scientific psychology after 100 years: The unconscious mind in the era of cognitive neuroscience. *Philosophy, Psychiatry, and Psychology* 2:123-34.

Zilboorg, G. 1951. Variations in the scope of awareness. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the Second Conference_*. Josiah Macy Foundation.

6.2n Consciousness and Time

Allport, D.A. 1968. Phenomenal similarity and the perceptual moment hypothesis. *British Journal of Psychology* 59:395-406.

Banks, R. & Cappon, D. 1962. Effect of reduced sensory input on time perception. *Perceptual and Motor Skills* 14:74.

Block, R.A. 1979. Time and consciousness. In (G. Underwood & R. Stevens, eds) *_Aspects of Consciousness, Volume 1_*. Academic Press.

Block, R.A. (ed) 1990. *_Cognitive Models of Psychological Time_*. Lawrence Erlbaum.

- Block, R.A. 1996. Psychological time and memory systems of the brain. In (J. Fraser & M. Soulsby, eds) *_Dimensions of Time and Life: The Study of Time_*, volume 8. International Universities Press.
- Brown, J.W. 1990. Psychology of time awareness. *Brain and Cognition* 14:144-64.
- Brown, J.W. 1991. *_Self and Process: Brain States and the Conscious Present_*. Springer-Verlag.
- Brown, J. 2000. *_Mind and Nature: Essays on Time and Subjectivity_*. Whurr Publishers.
- Cohen, J. 1954. The experience of time. *Acta Psychologica* 10:207-19.
- Dennett, D.C. & Kinsbourne, M. 1992. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15:183-201.
- Eisler, H. 1975. Subjective duration and psychophysics. *Psychological Review* 82:429-50.
- Fraser, J.T. (ed) 1989. *_Time and Mind: Interdisciplinary Issues_*. International Universities Press.
- Gooddy, W. 1967. Introduction to problems of time awareness. *Studium Generale* 20:33-41.
- Hicks, R.E., Miller, G.W., Gaes, G., & Bierman, K. 1977. Concurrent processing demands and the experience of time-in-passing. *American Journal of Psychology* 90:431-46.
- Hoagland, H. 1950. Consciousness and the chemistry of time. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the First Conference_*. Josiah Macy Foundation.
- Hoagland, H. 1943. The chemistry of time. *Scientific Monthly* 56:56-61.
- Knight, R.T. & Grabowecky, M. 1995. Escape from linear time: Prefrontal cortex and conscious experience. In (M. Gazzaniga, ed) *_The Cognitive Neurosciences_*. MIT Press.
- Melges, F.T. 1989. Disorders of time and the brain in severe mental illness. In (J. Fraser, ed) *_Time and Mind: Interdisciplinary Issues_*. International Universities Press.
- Michon, J.A. 1975. Time experience and memory processes. In (J. Fraser & N. Lawrence, eds) *_The Study of Time II_*. Springer-Verlag.
- Michon, J.A. 1972. Processing of temporal information and the cognitive theory of time experience. In (J. Fraser, F. Haber, & G. Muller, eds) *_The Study of Time_*. Springer-Verlag.
- Moiseeva, N.I. 19xx. Perception of time by human consciousness. *Chronobiologia* 15:301-317.
- Natsoulas, T. 1993. The stream of consciousness: II. William James's specious present. *Imagination, Cognition and Personality* 12:367-385.
- Newman, M.A. 1982. Time as an index of expanding consciousness with age. *Nursing Research* 31:290-293.
- Orme, J.E. 1969. *_Time, Experience and Behaviour_*. Illife.

Ornstein, R.E. 1969. *On the Experience of Time*. Harmondsworth.

Poppel, E. 1988. *Mindworks: Time and Conscious Experience*. Harcourt Brace Jovanovich.

Poppel, E. & Schwender, D. 1993. Temporal mechanisms of consciousness. *International Anesthesiology Clinics* 31:27-38.

Proust, J. 1994. Time and conscious experience. In (C.C. Gould, ed) *Artifacts, Representations, and Social Practice*. Kluwer.

Reidhead, V.A. & Wolford, J.B. 1998. Context, conditioning, and meaning of time-consciousness in a Trappist monastery. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Robertson, T.B. 1923. Consciousness and the sense of time. *Scientific Monthly* 16:649-657.

Sanders, S.A. 1986. Development of a tool to measure subjective time experience. *Nursing Research* 35:178-182.

Schaltenbrand, G. 1967. Consciousness and time. *Annals of the New York Academy of Sciences* 138:632-645.

Strong, C.A. 1896. Consciousness and time. *Psychological Review* 3:149-57.

Stroud, J.M. 1967. The fine structure of psychological time. *Annals of the New York Academy of Sciences* 138:623-631.

Stroud, J.M. 1957. The fine structure of psychological time. In (H. Quastler, ed) *Information Theory in Psychology: Problems and Methods*. Free Press.

6.2o Self-Consciousness

Asendorpf, J.B., Warkentin, V., & Baudonniere, P. 1996. Self-awareness and other-awareness. II: Mirror self-recognition, social contingency awareness, and synchronic imitation. *Developmental Psychology* 32q:313-321.

Butterworth, G. 1995. The self as an object of consciousness in infancy. In (P. Rochat, ed) *The Self in Infancy: Theory and Research*. Elsevier.

Carver, C. & Scheier, M.F. 1983. Self-awareness and the self-regulation of behaviour. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Cheeks, J.M. & Briggs, S.R. 1982. Self-consciousness and aspects of personality. *Journal of Research in Personality* 16:401-8.

Cooney, B. 1979. The neural basis of self-consciousness. *Nature and System* 1:16-31.

Duval, S. & Wicklund, R.A. 1972. *A Theory of Objective Self-Awareness*. Academic Press.

Dymond, S. & Barnes, D. 1997. Behavior-analytic approaches to self-awareness. *Psychological Record* 47:181-200.

Feinberg, T.E. 1997. Some interesting perturbations of the self in neurology. *Seminars in Neurology* 17:129-35.

Fenigstein, A., Scheier, M.F. & Buss, A.H. 1975. Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology* 43:522-27.

- Fenigstein, A. 1997. Self-consciousness and its relation to psychological mindedness. In (M. McCallum & W. Piper, eds) *Psychological Mindedness: A Contemporary Understanding*. Lawrence Erlbaum.
- Frith, C.D. 1996. The role of the prefrontal cortex in self-consciousness: the case of auditory hallucinations. *Philosophical Transactions of the Royal Society of London* B351:1505-12.
- Freeman, W. & Watts, J.W. 1941. The frontal lobes and consciousness of self. *Psychosomatic Medicine* 3:111-19.
- Gallup, G.G. 1998. Self-awareness and the evolution of social intelligence. *Behavioural Processes* 42:239-247.
- Gardiner, J. 2000. On the objectivity of subjective experiences and autonoetic and noetic consciousness. In (E. Tulving, ed) *Memory, Consciousness, and the Brain: The Tallinn Conference*. Psychology Press/Taylor & Francis.
- Hart, D. & Fegley, S. 1997. The development of self-awareness and self-understanding in cultural context. In (U. Neisser & D. Jopling, eds) *The Conceptual Self in Context*. Cambridge University Press.
- Hart, D. & Fegley, S. 1994. Social imitation and the emergence of a mental model of self. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.
- James, W. 1890. The consciousness of self. In *The Principles of Psychology*.
- Jaynes, J. 1976. *The Origins of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin.
- Kessel, F.S, Cole, P.M. & Johnson, D.L. (eds) *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum.
- Kihlstrom, J.F. 1997. Consciousness and me-ness. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.
- Kihlstrom, J.F. & Klein, S.B. 1997. Self-knowledge and self-awareness. In (J. Snodgrass, R. Thompson, eds) *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.
- Kinsbourne, M. 1995. Awareness of one's own body: An attentional theory of its nature, development, and brain basis. In (J. Bermudez, A. Marcel, & N. Eilan, eds) *The Body and the Self*. MIT Press.
- Kinsbourne, M. 1998. Representations in consciousness and the neuropsychology of insight. In (X. Amador & A. David, eds) *Insight and Psychosis*. Oxford University Press.
- Kunzendorf, R.G. 1988. Self-consciousness as the monitoring of cognitive states: A theoretical perspective. *Imagination, Cognition and Personality* 7:3-22.
- Kunzendorf, R.G., Beltz, S.M. & Tymowicz, G. 1992. Self-awareness in autistic subjects and deeply hypnotized subjects: Dissociation of self-concept versus self-consciousness. *Imagination, Cognition and Personality* 11:129-41.
- Kunzendorf, R. 2000. Individual differences in self-conscious source monitoring: Theoretical, experimental, and clinical considerations. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

- Levine, B. 2000. Self-regulation and autonoetic consciousness. In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.
- Lewis, M. 1991. Ways of knowing: Objective self-awareness or consciousness. *Developmental Review* 11:231-43.
- Lewis, M. 1994. Myself and me. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- Markova, I. 1990. The development of self-consciousness: Baldwin, Mead, and Vygotsky. In (J. Faulconer & R. Williams, eds) *_Reconsidering Psychology_*. Duquesne University Press.
- Mitchell, R.W. 1993. Mental models of mirror self-recognition: Two theories. *New Ideas in Psychology* 11:295-325.
- Mitchell, R.W. 1994. Multiplicities of self. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- Mitchell, R.W. 1997. A comparison of the self-awareness and kinesthetic-visual matching theories of self-recognition: Autistic children and others. In (J. Snodgrass, R. Thompson, eds) *_The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept_*. New York Academy of Sciences.
- Mollon, P. 1987. Self-awareness, self-consciousness, and preoccupation with self. In (K. Yardley & T. Honess, eds) *_Self and Identity: Psychosocial Perspectives_*. Wiley.
- Morin, A. & Everett, J. 1990. Inner speech as a mediator of self-awareness, self-consciousness, and self-knowledge: An hypothesis. *New Ideas in Psychology* 8:337-56.
- Nasby, W. 1989. Private self-consciousness, self-awareness, and the reliability of self-reports. *Journal of Personality and Social Psychology* 56:950-7.
- Neisser, U. 1991. Five kinds of self-knowledge. *Philosophical Psychology* 1:35-59.
- Neisser, U. 1992. The development of consciousness and the acquisition of self. In (F. Kessel, P. Cole, & D.L. Johnson, eds) *_Self and Consciousness: Multiple Perspectives_*. Lawrence Erlbaum.
- Parker, S.T., Mitchell, R.M., & Boccia, M.L. 1994. *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- Sass, L. 2000. Schizophrenia, self-experience, and the so-called "negative symptoms": Reflections on hyperreflexivity. In (D. Zahavi, ed) *_Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience_*. John Benjamins.
- Shotter, J. 1983. Consciousness and self-consciousness: Inner games and alternative realities. In (G. Underwood, ed) *_Aspects of Consciousness, Volume 3: Awareness and Self-Awareness_*. Academic Press.
- Shrauger, J.S. & Osberg, T.M. 1983. Self-awareness: The ability to predict one's subsequent behaviour. In (G. Underwood, ed) *_Aspects of Consciousness, Volume 3: Awareness and Self-Awareness_*. Academic Press.

Siegrist, M. 1995. Inner speech as a cognitive process mediating self-consciousness and inhibiting self-deception. *Psychological Reports* 76:259-65.

Snodgrass, J.G. & Thompson, R.L. (eds) 1997. *The Self across Psychology: Self-recognition, Self-awareness, and the Self Concept*. New York Academy of Sciences.

Titchener, E.B. 1911. A note on the consciousness of self. *American Journal of Psychology* 22:540-52.

Watson, J.S. 1994. Detection of self: The perfect algorithm. In (S. Parker, R. Mitchell, & M. Boccia, eds) *Self-Awareness in Animals and Humans: Developmental Perspectives*. Cambridge University Press.

Watson, P.J., Morris, R.J., Ramsey, A. Hickman, S.E. 1996. Further contrasts between self-reflectiveness and internal state awareness factors of private self-consciousness. *Journal of Psychology* 130:183-92.

6.2p Development of Consciousness

Anderson, J.R. 1984. The development of self-recognition: A review. *Developmental Psychobiology* 17:35-49.

Brainerd, C.J., Stein, L.M., & Reyna, V.F. 1998. On the development of conscious and unconscious memory. *Developmental Psychology* 34:342-357.

Briskin, A.S. 1974. A developmental model of self-awareness. *Counseling and Values* 18:79-85.

Burgess, J.A. & Tawia, S.A. 1996. When did you first begin to feel it? Locating the beginnings of human consciousness? *Bioethics* 10:1-26.

Butterworth, G. 1995. The self as an object of consciousness in infancy. In (P. Rochat, ed) *The Self in Infancy: Theory and Research*. Elsevier.

Davis, L.H. 1989. Self-consciousness in chimps and pigeons. *Philosophical Psychology* 2:249-59.

Flavell, J.H. 1993. Young children's understanding of thinking and consciousness. *Current Directions in Psychological Science* 2:40-43.

Flavell, J.H., Green, F.L., & Flavell, E.R. 1993. Children's understanding of the stream of consciousness. *Child Development* 64:387-398.

Flavell, J.H., Green, F.L., Flavell, E.R. & Grossman, J.B. 1997. The development of children's knowledge about inner speech. *Child Development* 68:39-47.

Flavell, J.H., Green, F.L., Flavell, E.R. 1995. The development of children's knowledge about attentional focus. *Developmental Psychology* 31:706-12.

Flavell, J., Green, F., & Flavell, E. 2000. Development of children's awareness of their own thoughts. *Journal of Cognition & Development* 1:97-112.

Foulkes, D. 1999. *Children's Dreaming and the Development of Consciousness*. Harvard University Press.

Gallagher, S. & Meltzoff, A. 1996. The earliest sense of self and others: Merleau-Ponty and recent developmental studies. *Philosophical Psychology* 9:211-33.

Gopnik, A. & Meltzoff, A.N. 1994. Minds, bodies, and persons: Young children's

understanding of the self and others as reflected in imitation and theory of mind research. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.

Griffin, S. 1991. Young children's awareness of their inner world: A neo-structural analysis of the development of intrapersonal intelligence. In (R. Case, ed) *_The Mind's Staircase: Exploring the Conceptual Underpinnings of Children's Thought and Knowledge_*. Lawrence Erlbaum.

Kagan, J. 1981. *_The Second Year: The Emergence of Self-Awareness_*. Harvard University Press.

Kuhn, D. 2000. Metacognitive development. *Current Directions in Psychological Science* 9:178-181.

Lewis, M. 1990. The development of intentionality and the role of consciousness. *Psychological Inquiry* 1:231-247.

Lewis, M. 1991. Ways of knowing: Objective self-awareness or consciousness. *Developmental Review* 11:231-43.

Lunzer, E.A. 1979. The development of consciousness. In (G. Underwood & R. Stevens, eds) *_Aspects of Consciousness_*. Academic Press.

Marbach, E. 1987. Laws of consciousness as norms of mental development. In (B. Inhelder, D. de Caprona, & A. Cornu-Wells, eds) *_Piaget Today_*. Lawrence Erlbaum.

Markova, I. 1990. The development of self-consciousness: Baldwin, Mead, and Vygotsky. In (J. Faulconer & R. Williams, eds) *_Reconsidering Psychology_*. Duquesne University Press.

McCune, L. 1993. The development of play as the development of consciousness. In (M. Bornstein & A. O'Reilly, eds) *_The Role of Play in the Development of Thought_*. Jossey-Bass.

Mounoud, P. 1990. Consciousness as a necessary transitional phenomenon in cognitive development. *Psychological Inquiry* 1:253-58.

Neisser, U. 1992. The development of consciousness and the acquisition of skill. In (F. Kessel, P. Cole, & D. Johnson, eds) *_Self and Consciousness: Multiple Perspectives_*. Lawrence Erlbaum.

Parker, S.T., Mitchell, R.M., & Boccia, M.L. 1994. *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.

Piaget, J. 1954. The problem of consciousness in child psychology: Developmental changes in awareness. In (H. Abramson, ed) *_Problems of Consciousness: Transactions of the Fourth Conference_*. Josiah Macy Foundation.

van Eenwyk, J.R. 1996. Chaotic dynamics and the development of consciousness. In (E. MacCormac & M. Stamenov, eds) *_Fractals of Brain, Fractals of Mind: In Search of a Symmetry Bond_*. John Benjamins.

Wheeler, M. 2000. Varieties of consciousness and memory in the developing child. In (E. Tulving, ed) *_Memory, Consciousness, and the Brain: The Tallinn Conference_*. Psychology Press/Taylor & Francis.

Wilber, K. 1979. A developmental view of consciousness. *Journal of Transpersonal Psychology* 11:1-21.

Zelazo, P. D. 1996. Towards a characterization of minimal consciousness. *New Ideas in Psychology* 14:63-80.

Zelazo, P.R. & Zelazo, P.D. 1998. The emergence of consciousness. In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Zelazo, P. 2000. Self-reflection and the development of consciously controlled processing. In (P. Mitchell & K. Riggs, eds) *_Children's Reasoning and the Mind_*. Psychology Press/Taylor & Francis.

6.2q The Stream of Consciousness

Antrobus, J.S., Singer, J.L., & Greenberg, S. 1966. Studies in the stream of consciousness: Experimental enhancement and suppression of spontaneous cognitive processes. *Perceptual and Motor Skills* 23:399-417.

Baars, B.J. 1993. How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity? In *_Experimental and Theoretical Studies of Consciousness_* (Ciba Foundation Symposium 174).

Bakan, P. 1978. Two streams of consciousness: A typological approach. In (K. Pope & J. Singer, eds) *_The Stream of Consciousness: Scientific Investigation into the Flow of Experience_*. Plenum.

Bonanno, G.A. & Singer, J.L. 1993. Controlling one's stream of thought through perceptual and reflective processing. In (D. Wegner & J. Pennebaker, eds) *_Handbook of Mental Control_*. Prentice-Hall.

Capek, M. 1950. Stream of consciousness and "duree reelle." *Philosophy and Phenomenological Research* 10:331-353.

Diaz, J. 1996. The stream revisited: A process model of phenomenological consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.

Dainton, B. 2000. *_Stream of Consciousness: Unity and continuity in conscious experience_*. Routledge.

Flanagan, O. 1992. The stream of consciousness. In *_Consciousness Reconsidered_*. MIT Press.

Flavell, J.H., Green, F.L., & Flavell, E.R. 1993. Children's understanding of the stream of consciousness. *Child Development* 64:387-398.

Gurwitsch, A. 1943. William James' theory of the "transitive parts" of the stream of consciousness. *Philosophy and Phenomenological Research* 3:449-477.

James, W. 1990. The stream of thought. In *_Principles of Psychology_*.

James, W. 1992. The stream of consciousness. In *_Psychology: The Briefer Course_*.

Klinger, E. 1978. Modes of normal conscious flow. In (K. Pope & J. Singer, eds) *_The Stream of Consciousness: Scientific Investigation into the Flow of Experience_*. Plenum.

Mueller, E.T. 1990. *_Daydreaming in Humans and Machines: A Computer Model of the Stream of Thought_*. Ablex.

Natsoulas, T. 1987. The six basic concepts of consciousness and William James' stream of thought. *Imagination, Cognition, and Personality* 6:289-319.

Natsoulas, T. 1988. Sympathy, empathy, and the stream of consciousness. *Journal for the Theory of Social Behaviour* 18:169-195.

Natsoulas, T. 1992-1996. The stream of consciousness: Parts I-XVI. *Imagination, Cognition, and Personality* 12:3-21, 12:367-85, 13:73-90, 13:229-47, 13:347-66, 14:59-77, 14:131-49, 14:333-52, 15:171-91, 16:161-80, 16:281-300, 17:45-64, 17:123-40, 17:229-47.

Penfield, W. 1955. The permanent record of the stream of consciousness. *Acta Psychologica* 11:47-69.

Petchkovsky, L. 2000. 'Stream of consciousness' and 'ownership of thought' in indigenous people in Central Australia. *Journal of Analytical Psychology* 45:577-597.

Pollio, H.R. 1990. The stream of consciousness since James. In (M. Johnson & T. Henley, eds) *_Reflections on "The Principles of Psychology": William James after a Century_*. Lawrence Erlbaum.

Pope, K.S. & Singer, J.L. (eds) 1978. *_The Stream of Consciousness: Scientific Investigations into the Flow of Human Experience_*. Plenum Press.

Pope, K.S. 1978. How gender, solitude, and posture influence the stream of consciousness. In (K. Pope & J. Singer, eds) *_The Stream of Consciousness: Scientific Investigation into the Flow of Experience_*. Plenum.

Rychlak, J.F. 1978. The stream of consciousness: Implications for a humanistic psychological theory. In (K. Pope & J. Singer, eds) *_The Stream of Consciousness: Scientific Investigation into the Flow of Experience_*. Plenum.

Schuetz, A. 1940. William James's concept of the stream of thought, phenomenologically interpreted. *Journal of Philosophy* 37:673-74.

Singer, J.L. 1974. Daydreaming and the stream of thought. *American Scientist* 62:417-425.

Singer, J.L. 1975. Navigating the stream of consciousness: Research in daydreaming and related inner experience. *American Psychologist* 30:727-738.

Singer, J.L. 1978. Experimental studies of daydreaming and the stream of thought. In (K. Pope & J. Singer, eds) *_The Stream of Consciousness: Scientific Investigation into the Flow of Experience_*. Plenum.

Singer, J.L. 1998. Daydreams, the stream of consciousness, and self-representations. In (R. Bornstein & J. Masling, eds) *_Empirical Perspectives on the Psychoanalytic Unconscious_*. American Psychological Association.

Strange, J.R. 1978. A search for the sources of the stream of consciousness. In (K. Pope & J. Singer, eds) *_The Stream of Consciousness: Scientific Investigation into the Flow of Experience_*. Plenum.

6.2r Foundational Issues

Baars, B.J. 1986. What is a theory of consciousness a theory of? The search for criterial constraints on theory. *Imagination, Cognition, and Personality* 1:3-24.

Baars, B.J. 1994. A thoroughly empirical approach to consciousness. *Psyche* 1.

Baars, B.J. 1996. Understanding subjectivity: Global workspace theory and

the resurrection of the observing self. *Journal of Consciousness Studies* 3:211-17.

Benoit, P.J. & Benoit, W.L. 1986. Consciousness: The mindlessness/mindfulness and verbal report controversies. *Western Journal of Speech Communication* 50:41-63.

Bindra, D. 1970. The problem of subjective experience. *Psychological Review* 77:581-84.

Blanshard, B. & Skinner, B.F. 1966. The problem of consciousness: A debate. *Philosophy and Phenomenological Research* 27:317-37.

Carlson, R.A. 1992. Starting with consciousness. *American Journal of Psychology* 105:598-604.

Casler L. 1976. The "consciousness problem" is not the problem. *Perceptual and Motor Skills* 42:227-32.

Hebb, D.O. 1954. The problem of consciousness and introspection. In (J. Delafresnaye, ed) *_Brain Mechanisms and Consciousness_*. Blackwell.

Kihlstrom, J.F. 1987. What this discipline needs is a good ten-cent taxonomy of consciousness. *Canadian Psychology* 28:116-118.

Merikle, P.M. 1984. Toward a definition of awareness. *Bulletin of the Psychonomic Society* 22:449-50.

Natsoulas, T. 1974. The subjective, experiential element in perception. *Psychological Bulletin* 81:611-31.

Natsoulas, T. 1981. Basic problems of consciousness. *Journal of Personality and Social Psychology* 41:132-78.

Natsoulas, T. 1990. Is consciousness what psychologists actually examine? *American Journal of Psychology* 105:363-84.

Revonsuo, A. 1993. Is there a ghost in the cognitive machinery? *Philosophical Psychology* 6:387-405.

Rychlak, J.F. 1997. *_In Defense of Human Consciousness_*. American Psychological Association.

Tolman, E.C. 1935. Psychology versus immediate experience. *Philosophy of Science* 2:356-80.

Wilson, D.L. 1978. Brain mechanisms, consciousness, and introspection. In (A. Sugarman & R. Tarter, eds) *_Expanding Dimensions of Consciousness_*. Springer.

Zener, K. 1952. Significance of the experience of the individual for the science of psychology. *Minnesota Studies in the Philosophy of Science* 2:354-69.

6.2s Consciousness and Psychology, Misc

Aurell, G. 1979. Perception: A model comprising two modes of consciousness. *Perceptual and Motor Skills* 49:431-44.

Aurell, G. 1989. Man's triune conscious mind, parts I, II, and III. *Perceptual and Motor Skills* 68:747-54, 78:31-39, 81:463-66

Bolton, N. 1983. Forms of awareness. In (G. Underwood, ed) *_Aspects of*

Consciousness, Volume 3: Awareness and Self-Awareness_. Academic Press.

Bowden, E.M. 1997. The effect of reportable and unreportable hints on anagram solution and the aha! experience. *Consciousness and Cognition* 6:545-573.

Davidson, J.M. & Davidson, R.J. (eds) 1980. *The Psychobiology of Consciousness*. Plenum.

Davidson, R., Schwartz, G. & Shapiro, D. (eds) 1983. *Consciousness and Self-Regulation*. Plenum.

Delacour, J. 1997. Object perception and recognition: A model for the scientific study of consciousness. *Theory and Psychology* 7:257-62.

Dulany, D.E. 1991. Conscious representation and thought systems. In (R. Wyer & T. Srull, eds) *The Content, Structure, and Operation of Thought Systems*. Lawrence Erlbaum.

Dulany, D.E. 1997. Consciousness in the explicit (deliberative) and implicit (evocative). In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Farthing, G.W. 1992. *The Psychology of Consciousness*. Prentice Hall.

Frith C.D. 1979. Consciousness, information processing and schizophrenia. *British Journal of Psychiatry* 134:225-35.

Greenberg, G. & Tobach, E. (eds) 1987. *Cognition, language, and consciousness: Integrative levels*. Lawrence Erlbaum.

Higgins, E.E. & Bargh, J.A. 1992. Unconscious sources of subjectivity and suffering: Is consciousness the solution? In (L. Martin & A. Tesser, eds) *The Construction of Social Judgments*. Lawrence Erlbaum

Hilgard, E.R. 1977. *Divided Consciousness: Multiple Controls in Human Thought and Action*. Wiley.

Hilgard, E.R. 1977. Controversies over consciousness and the rise of cognitive psychology. *Australian Psychologist* 12:7-27.

Hilgard E.R. 1977. The problem of divided consciousness: A neodissociation interpretation. *Annals of the New York Academy of Sciences* 296:48-59.

Hilgard, E.R. 1980. Consciousness in contemporary psychology. *Annual Review of Psychology* 31:1-26.

Hilgard, E.R. 1992. Divided consciousness and dissociation. *Consciousness and Cognition* 1:16-31.

Hirst, W. 1995. Cognitive aspects of consciousness. In (M. Gazzaniga, ed) *The Cognitive Neurosciences*. MIT Press.

Humphrey, N. 1992. *A History of the Mind*. Simon and Schuster.

Jonassen, D.H. 1979. Video-mediated, objective self-awareness, self-perception, and locus of control. *Perceptual and Motor Skills* 48:255-265.

Kihlstrom, J.F. 1993. The continuum of consciousness. *Consciousness and Cognition* 2:334-54.

Lashley, K.S. 1923. The behavioristic interpretation of consciousness. *Psychological Review* 30:237-72.

Lewicki, P., Czyzewska, M. & Hill, T. 1997. Cognitive mechanisms for acquiring

- "experience": The dissociation between conscious and nonconscious cognition. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.
- Mandler, G. 1975. Consciousness: respectable, useful, and probably necessary. In (R. Solso, ed) *_Information Processing and Cognition_*. Lawrence Erlbaum.
- Mandler, G. 1988. Problems and direction in the study of consciousness. In (M. Horowitz, ed) *_Psychodynamics and Cognition_*. University of Chicago Press.
- Mandler, G. 1992. Toward a theory of consciousness. In (H.G. Geissler, S.W. Link, & J.T. Townsend, eds) *_Cognition, Information Processing, and Psychophysics: Basic Issues_*. Lawrence Erlbaum.
- Mandler, G. 1997. Consciousness redux. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.
- Mandler, G. & Nakamura, Y. 1987. Aspects of consciousness. *Personality and Social Psychology Bulletin* 13:299-313.
- Natsoulas, T. 1984. Personality and consciousness: A theoretical essay. *Cognition and Brain Theory* 7:135-66.
- Oakley, D.A. & Eames, L.C. 1986. The plurality of consciousness. In (D. Oakley, ed) *_Mind and Brain_*. Methuen.
- Oatley, K. 1988. On changing one's mind: A possible function of consciousness. In (A. Marcel & E. Bisiach, eds) *_Consciousness in Contemporary Science_*. Oxford University Press.
- Ornstein, R.E. (ed) 1974. *_The Nature of Human Consciousness: A Book of Readings_*. Viking Press.
- Ornstein, R.E. 1977. *_The Psychology of Consciousness_*. Harcourt Brace Jovanovich.
- Posner, M.I. & Klein, M. 1973. On the functions of consciousness. In (S. Kornblum, ed) *_Attention and Performance_*, vol 4. Academic Press.
- Schwartz, G. & Shapiro, D. (eds) 1976. *_Consciousness and Self-regulation_*. Plenum.
- Schwartz, G. 2000. Individual differences in subtle awareness and levels of awareness: Olfaction as a model system. In (R. Kunzendorf & B. Wallace, eds) *_Individual Differences in Conscious Experience_*. John Benjamins.
- Schwarz, N. & Clore, G.L. 1996. Feelings and phenomenal experiences. In *_Social Psychology: Handbook of Basic Principles_*. Guilford Press.
- Shallice, T. 1991. The revival of consciousness in cognitive science. In (W. Kessen, A. Ortony, & F. Craik, eds) *_Memories, Thoughts, and Emotions: Essays in Honor of George Mandler_*. Lawrence Erlbaum.
- Solso, R. (ed) 1975. *_Information Processing and Consciousness_*. Lawrence Erlbaum.
- Sperry, R.W. 1987. Structure and significance of the consciousness revolution. *Journal of Mind & Behavior* 8:37-65.
- Sperry, R.W. 1995. The riddle of consciousness and the changing scientific worldview. *Journal of Humanistic Psychology* 35:7-33.
- Stout, M. 2001. *_The Myth of Sanity: Divided Consciousness and the Promise of Awareness_*. Viking/Penguin Books.

Underwood, G. & Stevens, R. (eds) 1979. Aspects of Consciousness: Volume 1, Psychological Issues. Academic Press.

Underwood, G. & Stevens, R. (eds) 1981. Aspects of Consciousness: Volume 2, Structural Issues. Academic Press.

Underwood, G. & Stevens, R. (eds) 1982. Aspects of Consciousness: Volume 3, Awareness and Self-Awareness. Academic Press.

Underwood, G. & Stevens, R. (eds) 1984. Aspects of Consciousness: Volume 4, Clinical Issues. Academic Press.

Underwood, G. (ed) 1996. Implicit Cognition. Oxford University Press.

von der Malsburg, C. 1986. Am I thinking assemblies? In (G. Palm & A. Aertsen, eds) Brain Theory. Springer.

von der Malsburg, C. 1997. The coherence definition of consciousness. In (M. Ito, Y. Miyashita, & E.T. Rolls, eds) Cognition, Computation, and Consciousness. Oxford University Press.

Wallace, B. & Fisher, L. 2000. Biological rhythms and individual differences in consciousness. In (R. Kunzendorf & B. Wallace, eds) Individual Differences in Conscious Experience. John Benjamins.

Wilks, Y. 1984. Machines and consciousness. In (C. Hookway, ed) Minds, Machines and Evolution. Cambridge University Press.

6.3 Consciousness and Physics

6.3a The Interpretation of Quantum Mechanics

Albert, D. & Loewer, A. 1988. Interpreting the many-worlds interpretation. *Synthese* 77:195-213.

Butterfield, J. 1996. Whither the minds? *British Journal for the Philosophy of Science* 47:200-??.

Butterfield, J. 1998. Quantum curiosities of psychophysics. In (J. Cornwell, ed) Consciousness and Human Identity. Oxford University Press.

Byrne, A. & Hall, N. 1999. Chalmers on consciousness and quantum mechanics. *Philosophy of Science* 66:370-90.

Goertzel, B. 1992. Quantum theory and consciousness. *Journal of Mind and Behavior* 13:29-36.

Goswami, A. 1989. The idealistic interpretation of quantum mechanics. *Physics Essays* 2:385-400.

Goswami, A. 1990. Consciousness in quantum physics and the mind-body problem. *Journal of Mind and Behavior* 11:75-96.

Klein, S. 1991. The duality of psycho-physics. In (A. Gorea, ed) Representations of Vision. Cambridge University Press.

Lehner, C. 1997. What it feels like to be in a superposition, and why: Consciousness and the interpretation of Everett's quantum mechanics. *Synthese* 110:191-216.

Lockwood, M. 1989. Mind, Brain, and the Quantum. Oxford University Press.

- Lockwood, M. 1996. Many-minds interpretations of quantum mechanics. *British Journal for the Philosophy of Science* 47:159-88.
- Mulhauser, G. 1995. Materialism and the "problem" of quantum measurement. *Minds and Machines* 5:207-17.
- Mulhauser, G. 1995. On the end of a quantum-mechanical romance. *Psyche* 2(19).
- Page, D.N. 1995. Attaching theories of consciousness to Bohmian quantum mechanics. Manuscript.
- Page, D.N. 1996. Sensible quantum mechanics: Are probabilities only in the mind? *International Journal of Modern Physics D* 5:583-96.
- Penrose, R. 1987. Quantum physics and conscious thought. In (B. Hiley & D. Peat, eds) *Quantum Implications: Essays in Honour of David Bohm*. Methuen.
- Shanks, N. 1995. Minds, brains, and quantum mechanics. *Southern Journal of Philosophy* 33:243-60.
- Squires, E.J. 1991. One mind or many? A note on the Everett interpretation of quantum theory. *Synthese* 89:283-6.
- Squires, E.J. 1993. Quantum theory and the relation between the conscious mind and the physical world. *Synthese* 97:109-23.
- Squires, E.J. 1994. Quantum theory and the need for consciousness. *Journal of Consciousness Studies* 1:201-4.
- Squires, E.J. 1998. Why are quantum theorists interested in consciousness? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Stapp, H.P. 1991. Quantum propensities and the brain-mind connection. *Foundations of Physics* 21:1451-77.
- Stapp, H.P. 1993. *Mind, Matter, and Quantum Mechanics*. Springer-Verlag.
- Stapp, H.P. 1995. Why classical mechanics cannot accommodate consciousness but quantum mechanics can. *Psyche* 2(5).
- Stapp, H.P. 1998. The evolution of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.
- Wigner, E. 1961. Remarks on the mind-body problem. In (I. Good, ed) *The Scientist Speculates*. Heineman.
- Woo, C.H. 1981. Consciousness and quantum interference: An experimental approach. *Foundations of Physics* 11:933-44.

6.3b Quantum Mechanisms of Consciousness

- Bass, L. 1975. A quantum-mechanical mind-body interaction. *Foundations of Physics* 5:159-72.
- Beck, F. & Eccles, J. 1992. Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Science USA* 89:11357-61.
- Beck, F. 1998. Synaptic transmission, quantum-state selection, and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

- Berezin, A.A. 1992. Correlated isotopic tunneling as a possible model for consciousness. *Journal of Theoretical Biology* 154:415-20.
- Dyer, M.G. 1994. Quantum physics and consciousness, creativity, computers: A commentary on Goswami's quantum-based theory of consciousness and free will. *Journal of Mind and Behavior* 15:265-90.
- Eccles, J.C. 1986. Do mental events cause neural events analogously to the probability fields of quantum mechanics? *Proceedings of the Royal Society of London B* 227:411-28.
- Germine, M. 1991. Consciousness and synchronicity. *Medical Hypotheses* 36:277-83.
- Globus, G. 1997. Nonlinear brain systems with nonlocal degrees of freedom. *Journal of Mind and Behavior*.
- Globus, G. 1998. Self, cognition, qualia, and world in quantum brain dynamics. *Journal of Consciousness Studies* 5:34-52.
- Grush, R. & Churchland, P. 1995. Gaps in Penrose's toiling. In (T. Metzinger, ed) *_Conscious Experience_*. Ferdinand Schoningh.
- Hameroff, S.R. 1994. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1:91-118.
- Hameroff, S.R. & Penrose, R. 1996. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness_*. MIT Press.
- Hameroff, S.R. & Penrose, R. 1996. Conscious events as orchestrated space-time selections. *Journal of Consciousness Studies* 3:36-53. Reprinted in (J. Shear, ed) *_Explaining Consciousness: The Hard Problem_*. MIT Press.
- Hameroff, S.R. & Scott, A. 1998. A Sonoran afternoon: A dialogue on quantum mechanics and consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.
- Jibu, M. & Yasue, K. 1995. *_Quantum Brain Dynamics and Consciousness: An Introduction_*. John Benjamins.
- Jibu, M. & Yasue, K. 1997. Magic without magic: Meaning of quantum brain dynamics. *Journal of Mind and Behavior*.
- King, C. 1997. Chaos, quantum mechanics, and the conscious brain. *Journal of Mind and Behavior*.
- Lahav, R. & Shanks, N. 1992. How to be a scientifically respectable 'property dualist'. *Journal of Mind and Behavior* 13:211-32.
- Marshall, I.N. 1989. Consciousness and Bose-Einstein condensates. *New Ideas in Psychology* 7:73-83.
- Marshall, I.N. 1995. Some phenomenological implications of a quantum model of consciousness. *Minds and Machines* 5:609-20.
- Penrose, R. 1994. Mechanisms, microtubules, and the mind. *Journal of Consciousness Studies* 1:241-49.
- Scott, A. 1996. On quantum theories of the mind. *Journal of Consciousness Studies* 3:484-91.
- Stapp, H.P. 1985. Consciousness and values in the quantum universe.

Foundations of Physics 15:35-47.

Stapp, H.P. 1994. Theoretical model of a purported empirical violation of the predictions of quantum mechanics. Physical Review A 50:18-22.

Stapp, H.P. 1995. The hard problem: A quantum approach. Journal of Consciousness Studies 3:194-210. Reprinted in (J. Shear, ed) Explaining Consciousness: The Hard Problem. MIT Press.

Stapp, H.P. 1997. Science of consciousness and the hard problem. Journal of Mind and Behavior 18:171-93.

Triffet, T. & Green, H.S. 1996. Consciousness: Computing the uncomputable. Mathematical and Computational Modelling 24:37-56.

Wolf, F.A. 1996. On the quantum mechanics of dreams and the emergence of self-awareness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness. MIT Press.

Zohar, D. 1995. A quantum-mechanical model of consciousness and the emergence of 'I'. Minds and Machines 5:597-607.

Zohar, D. 1996. Consciousness and Bose-Einstein condensates. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness. MIT Press.

6.3c Consciousness and Physics, Misc

Bilodeau, D. 1996. Physics, machines, and the hard problem. Journal of Consciousness Studies 3:386-401. Reprinted in (J. Shear, ed) Explaining Consciousness: The Hard Problem. MIT Press.

Bohm, D.J. 1986. A new theory of the relationship of mind and matter. Journal of the American Society for Psychical Research 80:113-35.

Burns, J. 1990. Contemporary models of consciousness, parts I & II. Journal of Mind and Behavior 11:153-171 & 12:407-420.

Clarke, C.J.S. 1995. The nonlocality of mind. Journal of Consciousness Studies 2:231-40. Reprinted in (J. Shear, ed) Explaining Consciousness: The Hard Problem. MIT Press.

Culbertson, J. 1982. Consciousness: Natural and Artificial. Libra.

de Silva, F. 1996. Consciousness and special relativity. IEEE Engineering in Medicine and Biology Magazine 15:21-26.

Dyer, M.G. 1994. Quantum physics and consciousness, creativity, and computers. Journal of Mind and Behavior 15:265-90.

Elitzur, A. 1996. Time and consciousness: The uneasy bearing of relativity on the mind-body problem. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness. MIT Press.

Herbert, N. 1993. Elemental Mind: Human Consciousness and the New Physics. Dutton.

Ho, M.W. 1997. Quantum coherence and conscious experience. Kybernetes 26:265-76.

Hodgson, D. 1988. The Mind Matters: Consciousness and Choice in a Quantum World. Oxford: Oxford University Press.

Hodgson, D. 1996. Nonlocality, local indeterminism, and consciousness. Ratio

9:1-22.

Nair, R. 1991. Quantum physics and the philosophy of mind: An essay review. *Journal of Scientific and Industrial Research* 50:66975.

Nunn, C.M.H., Clarke, C.J.S. & Blott, B.H. 1994. Collapse of a quantum field may affect brain function. *Journal of Consciousness Studies* 1:127-39.

Nunn, C.M.H. 1996. On the geometry of consciousness. *Journal of Consciousness Studies* 3:477-83.

Penrose, R. 1989. *_The Emperor's New Mind_*. Oxford University Press.

Penrose, R. 1994. *_Shadows of the Mind_*. Oxford University Press.

Penrose, R. 1997. *_The Large, the Small, and the Human Mind_*. Cambridge University Press.

Squires, E. 1990. *_Conscious Mind in the Physical World_*. Adam Hilger.

Zohar, D. & Marshall, I. 1990. *_The Quantum Self_*. Morrow.

6.4 Consciousness and Science, Misc

6.4a Evolution of Consciousness

Arhem, P. & Liljenstrom, H. 1997. On the coevolution of consciousness and cognition. *Journal of Theoretical Biology* 187:601-12.

Barlow, H.B. 1980. Nature's joke: A conjecture on the biological role of consciousness. In (B. Josephson & V. Ramachandran, eds) *_Consciousness and the Physical World_*. Pergamon Press.

Barlow, H.B. 1987. The biological role of consciousness. In (C. Blakemore & S. Greenfield, eds) *_Mindwaves_*. Blackwell.

Cairns-Smith, A.G. 1996. *_Evolving the Mind: On the Nature of Matter and the Origin of Consciousness_*. Cambridge University Press.

Coan, R.W. 1989. Alternative views on the evolution of consciousness. *Journal of Human Psychology* 29:167-99.

Cotterill, R. 2000. Did consciousness evolve from self-paced probing of the environment, and not from reflexes? *Brain and Mind* 1:283-298.

Crook, J.H. 1980. *_The Evolution of Human Consciousness_*. Oxford University Press.

Donald, M. 1995. The neurobiology of human consciousness: An evolutionary approach. *Neuropsychologia* 33:1087-1102.

Eccles, J.C. 1992. Evolution of consciousness. *Proceedings of the National Academy of Sciences USA* 89:7320-24.

Calvin, W.H. 1991. *_The Ascent of Mind: Ice Age Climates and the Evolution of Intelligence_*. Bantam Books.

Crook, J.H. 1980. *_The Evolution of Human Consciousness_*. Oxford University Press.

Dennett, D.C. 1986. Julian Jaynes' software archaeology. *Canadian Psychology* 27:149-54.

Dewart, L. 1989. *Evolution and Consciousness: The Role of Speech in the Origin and Development of Human Nature*. University of Toronto Press.

Glynn, I.M. 1993. The evolution of consciousness: William James' unresolved problem. *Biological Reviews of the Cambridge Philosophical Society* 68:599-616.

Hameroff, S.R. 1998. Did consciousness cause the Cambrian evolutionary explosion? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Humphrey, N. 1992. *A History of the Mind: Evolution and the Birth of Consciousness*. Simon and Schuster.

Jaynes, J. 1976. *The Origins of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin.

Jonker A. 1987. The origin of the human mind. A speculation on the emergence of language and human consciousness. *Acta Biotheoretica* 36:129-77.

King, J.E., Rumbaugh, D.M. & Savage-Rumbaugh, E.S. 1998. Evolution of intelligence, language, and other emergent processes for consciousness: A comparative perspective. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Lindahl, B.I.B. 1997. Consciousness and biological evolution. *Journal of Theoretical Biology* 187:613-29.

Nichols, S. & Grantham, T. 2000. Adaptive complexity and phenomenal consciousness. *Philosophy Of Science* 67:648-670.

Ornstein, R. 1991. *The Evolution of Consciousness: Of Darwin, Freud, and Cranial Fire: The Origins of the Way We Think*. Prentice-Hall.

Povinelli, D.J. 1987. Monkeys, apes, mirrors, minds: The evolution of self-awareness in primates. *Human Evolution* 2:493-507.

Pribram, K.H., Jerison, H.J., McGuiness, D., & Eccles, J.C. 1982. The evolution of consciousness: A symposium. In (J. Eccles, ed) *Mind and Brain*. Paragon House.

Reber, A.S. 1992. The cognitive unconscious: An evolutionary perspective. *Consciousness and Cognition* 1:93-133.

Reber, A.S. 1992. An evolutionary context for the cognitive unconscious. *Philosophical Psychology* 5:33-51.

Reber, A. & Allen, R. 2000. Individual differences in implicit learning: Implications for the evolution of consciousness. In (R. Kunzendorf & B. Wallace, eds) *Individual differences in conscious experience*. John Benjamins.

Rogers, L.J. 1995. Evolution and development of brain asymmetry, and its relevance to language, tool use and consciousness. *International Journal of Comparative Psychology* 8:1-15.

Roth, G. 2000. The evolution and ontogeny of consciousness. In (T. Metzinger, ed) *Neural Correlates of Consciousness*. MIT Press.

Towers, B. 1979. Consciousness and the brain: Evolutionary aspects. In *Brain and Mind* (Ciba Foundation Symposium 69). Elsevier.

Vandervert, L.R. 1995. Chaos theory and the evolution of consciousness and mind: A thermodynamic/holographic resolution to the mind-body problem. *New Ideas in Psychology* 13:107-27.

6.4b Consciousness and Language

Arbib, M.A. 1972. Consciousness: The secondary role of language. *Journal of Philosophy* 69.

Bailey, W. 1986. Consciousness and action/motion theories of communication. *Western Journal of Speech Communication* 50:74-86.

Blachowicz, J. 1997. The dialogue of the soul with itself. *Journal of Consciousness Studies* 4:485-508.

Carruthers, P. 1996. The involvement of language in conscious thinking. In *_Language, Thought, and Consciousness_*. Cambridge University Press.

Chafe, W.L. 1980. The deployment of consciousness in the construction of narrative. In (W. Chafe, ed) *_The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production_*. Ablex.

Chafe, W.L. 1994. *_Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing_*. University of Chicago Press.

Chafe, W.L. 1996. How consciousness shapes language. *Pragmatics and Cognition* 4:35-54.

Chafe, W. 2000. A linguist's perspective on William James and "The Stream of Thought." *Consciousness & Cognition* 9:618-628.

Chapman, S.B. & Ulatowska, H.K. 1997. Discourse in dementia: Considerations of Consciousness. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

de Beaugrande, R. 1997. The "conscious and unconscious mind" in the theoretical discourse of modern linguistics. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

Fludernik, M., & Sell, R.D. 1995. The fictions of language and the languages of fiction: The linguistic representation of speech and consciousness. *Journal of Pragmatics* 24:557.

Johnston, P.K. 1997. Battle within: Shakespeare's brain and the nature of human consciousness. *Journal of Consciousness Studies* 4:365-73.

Langacker, R.W. 1997. Consciousness, construal, and subjectivity. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

Lecours, A.R. 1998. Language contrivance on consciousness (and vice versa). In (H. Jasper, L. Descarries, V. Castellucci, & S. Rossignol, eds) *_Consciousness: At the Frontiers of Neuroscience_*. Lippincott-Raven.

Macphail, E. 2000. The search for a mental Rubicon. In (C. Heyes & L. Huber, eds) *_The Evolution of Cognition_*. MIT Press.

Markey, J.F. 1925. The place of language habits in a behavioristic explanation of consciousness. *Psychological Review* 32:384-401.

Pronko, N.H. 1987. Language with or without consciousness. In (G. Greenberg & E. Tobach, eds) *_Cognition, Language and Consciousness: Integrative Levels_*. Lawrence Erlbaum.

Ricciardelli, L.A. 1993. Two components of metalinguistic awareness: Control

of linguistic processing and analysis of linguistic knowledge. *Applied Psycholinguistics* 14:349-367.

Schooler, J.W & Fiore, S.M. 1997. Consciousness and the limits of language: You can't always say what you think or think what you say. In (J. Cohen & J. Schooler, eds) *_Scientific Approaches to Consciousness_*. Lawrence Erlbaum.

Sekhar, A.C. 1948. Language and consciousness. *Indian Journal of Psychology* 23:79-84.

Sinha, V. 1987. Symbolic language not a pre-requisite for self-awareness. *Psycho-Lingua* 17:115-121.

Stamenov, M.I. (ed) 1997. *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

Stamenov, M.I. 1997. Grammar, meaning, and consciousness: What sentence structure can tell us about the structure of consciousness. In (M. Stamenov, ed) *_Language Structure, Discourse, and the Access to Consciousness_*. John Benjamins.

6.4c Animal Consciousness

Allen, G.E. 1987. Materialism and reductionism in the study of animal consciousness. In (G. Greenberg, E. Tobach, eds) *_Cognition, Language, and Consciousness: Integrative Levels_*. Lawrence Erlbaum.

Bekoff, M. 1992. Scientific ideology, animal consciousness, and animal protection: A principled plea for unabashed common sense. *New Ideas in Psychology* 10:79-94.

Bradshaw, R.H. 1998. Consciousness in nonhuman animals: Adopting the precautionary principle. *Journal of Consciousness Studies* 5:108-14.

Burghardt, G. 1985. Animal awareness: Current perceptions and historical perspective. *American Psychologist* 40:905-919.

Carruthers, P. 1989. Brute experience. *Journal of Philosophy* 258-69.

Cheney, D.L. & Seyfarth, R.M. 1990. *_How Monkeys See the World: Inside the Mind of Another Species_*. University of Chicago Press.

Crook, J.H. 1983. On attributing consciousness to animals. *Nature* 303:11-14.

Dawkins, M.S. 1993. *_Through Our Eyes Only: The Search for Animal Consciousness_*.

Dennett, D.C. 1995. Animal consciousness: What matters and why? *Social Research* 62:691-710.

Eccles J.C. 1982. Animal consciousness and human self-consciousness. *Experientia* 38:1384-91.

Gallup, G.G. 1985. Do minds exist in species other than our own? *Neuroscience and Biobehavioral Reviews* 9:631-41.

Griffin, D.R. 1981. *_The Question of Animal Awareness: Evolutionary Continuity of Mental Experience_*. William Kaufmann.

Griffin, D.R. 1985. Animal consciousness. *Neuroscience and Biobehavioral Reviews* 9:615-22.

Griffin, D.R. 1992. *_Animal Minds_*. University of Chicago Press.

- Griffin, D.R. 1995. Windows on animal minds. *Consciousness and Cognition* 4:194-204.
- Heyes, C.M. 1987. Cognisance of consciousness in the study of animal knowledge. In (W. Callebaut & R. Pinxten, eds) *_Evolutionary Epistemology: A Multiparadigm Program_*. Reidel.
- Hughes, H. 2001. *_Sensory Exotica: A World Beyond Human Experience_*. MIT Press.
- Jolley, N. 1995. Sensation, intentionality, and animal consciousness. *Ratio* 8:128-42.
- Jolly, A. 1991. Conscious chimpanzees? A review of recent literature. In (C. Ristau, ed) *_Cognitive Ethology_*. Lawrence Erlbaum.
- Latto, R. 1986. The question of animal consciousness. *Psychological Record* 36:309-14.
- Oakley, D.A. 1985. Animal awareness, consciousness, and self-image. In (D. Oakley, ed) *_Brain and Mind_*. Methuen.
- Radner, D. & Radner, M. 1996. *_Animal Consciousness_*. Prometheus Books.
- Reiss, D. 1998. Cognition and communication in dolphins: A question of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.
- Ristau, C.A. 1983. Language, cognition, and awareness in animals? *Annals of the New York Academy of Sciences* 406:170-86.
- Roberts, H. 1968. Consciousness in animals and automata. *Psychological Reports* 22:1226-28.
- Rollin, B.E. 1986. Animal consciousness and scientific change. *New Ideas in Psychology* 4:141-52.
- Rollin, B.E. 1989. *_The Unheeded Cry: Animal Consciousness, Animal Pain, and Science_*. Oxford University Press.
- Rothschild, M. 1993. Thinking about animal consciousness. *Journal of Natural History* 27:509-12.
- Rushen, J.P. 1985. The scientific status of animal consciousness. *Applied Animal Behaviour Science* 13:387-390.
- Savage-Rumbaugh, E.S. & Rumbaugh, D. 1998. Perspectives on consciousness, language, and other emergent processes in apes and humans. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *_Toward a Science of Consciousness II_*. MIT Press.
- Tye, M. 1997. The problem of simple minds: Is there anything it's like to be a honeybee? *Philosophical Studies* 88:289-317.
- van Rooijen, J. 1981. Are feelings adaptations? The basis of modern applied animal ethology. *Applied Animal Ethology* 7:187-89.
- Weiskrantz, L. 1995. The problem of animal consciousness in relation to neuropsychology. *Behavioral Brain Research* 71:171-75.

6.4d Animal Self-Consciousness

- Byrne, R.W. & Whiten, A. 1988. *_Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans_*. Oxford University Press.
- Epstein, R., Lanza, R.P. & Skinner, B.F. 1981. "Self-awareness" in the pigeon. *Science* 212:695-96.
- Gallup, G.G. 1970. Chimpanzees: Self-recognition. *Science* 167:86-87.
- Gallup, G.G. 1975. Toward an operational definition of self-awareness. In (R. Tuttle, ed) *_Socioecology and the Psychology of Primates_*. Mouton.
- Gallup, G.G. 1977. Self-recognition in primates: A comparative approach to the bidirectional properties of consciousness. *American Psychologist* 32:329-38.
- Gallup, G.G. 1979. Self-recognition in chimpanzees and man: A developmental and comparative perspective. In (M. Lewis & M. Rosenblum, eds) *_Genesis of Behavior, Volume 2_*. Plenum Press.
- Gallup, G.G. 1982. Self-awareness and the emergence of mind in primates. *American Journal of Primatology* 2:237-48.
- Gallup, G.G. 1987. Self-awareness. In (G. Mitchell, ed) *_Comparative Primate Biology, Volume 2_*. Liss.
- Gallup, G.G. 1991. Toward a comparative psychology of self-awareness: Species limitations and cognitive consequences. In (G. Goethals & J. Strauss, eds) *_The Self: An Interdisciplinary Perspective_*. Springer-Verlag.
- Gallup, G.G. 1994. Self-recognition: Research strategies and experimental design. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- Hart, D. & Karmel, M.P. 1996. Self-awareness and self-knowledge in humans, apes, and monkeys. In (A. Russon, K. Bard, & S. Parkers, eds) *_Reaching into Thought: The Minds of the Great Apes_*. Cambridge University Press.
- Heyes, C.M. 1994. Reflections on self-recognition in primates. *Animal Behaviour* 47:909-19.
- Hyatt, C.W. & Hopkins, W. 1994. Self-awareness in bonobos and chimpanzees: A comparative perspective. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- Marten, K. & Psarakos, S. 1992. Using self-view television to distinguish between self-examination and social behavior in the bottlenose dolphin. *Consciousness and Cognition* 4:205-24.
- Marten, K. & Psarakos, S. 1994. Evidence for self-awareness in the bottlenose dolphin. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- Miles, H.L. 1994. Me Chantek: The development of self-awareness in a signing orangutan. In (S. Parker, R. Mitchell, & M. Boccia, eds) *_Self-Awareness in Animals and Humans: Developmental Perspectives_*. Cambridge University Press.
- Moynihan, M.H. 1997. Self-awareness, with specific references to Coleoid cephalopods. In (R. Mitchell, N. Thompson, & H. Miles, eds) *_Anthropomorphism, Anecdotes, and Animals_*. SUNY Press.
- Parker, S.T. 1991. A developmental approach to the origins of self-recognition in great apes. *Human Evolution* 6:435-49.

Patterson, F.G.P. & Cohn, R. 1994. Self-recognition and self-awareness in lowland gorillas. In (S. Parker, R. Mitchell, & M. Boccia, eds) Self-Awareness in Animals and Humans: Developmental Perspectives. Cambridge University Press.

Povinelli, D.J. 1987. Monkeys, apes, mirrors, minds: The evolution of self-awareness in primates. *Human Evolution* 2:493-507.

Suarez, S.D. & Gallup, G.G. 1981. Self-recognition in chimpanzee and orangutans, but not gorillas. *Journal of Human Evolution* 10:175-88.

Swartz, K.B. & Evans, S. 1991. Not all chimpanzees show self-recognition. *Primates* 32:483-96.

6.4e Altered States of Consciousness

Atkinson, R.P. & H. Earl. 1996. Enhanced vigilance in guided meditation: Implications of altered consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness. MIT Press.

Austin, J.H. 1998. Zen and the Brain: Toward an Understanding of Meditation and Consciousness. MIT Press.

Forman, R. (ed) 1990. The Problem of Pure Consciousness: Mysticism and Philosophy. Oxford University Press.

Forman, R. 1998. What does mysticism have to teach us about consciousness? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness 1996. MIT Press.

Goleman, D. 1976. Meditation and consciousness: An Asian approach to mental health. *American Journal of Psychotherapy* 30:41-54.

Hilgard, E.R. 1979. Consciousness and control: Lessons from hypnosis. *Australian Journal of Clinical & Experimental Hypnosis* 7:103-15.

Hunt H.T. 1985. Cognition and states of consciousness: the necessity for empirical study of ordinary and nonordinary consciousness for contemporary cognitive psychology. *Perceptual and Motor Skills* 60:239-82.

Katz, J.M. 1983. Altered states of consciousness and emotion. *Imagination, Cognition and Personality* 2:37-50.

Keen, E. 2000. Chemicals for the Mind: Psychopharmacology and Human Consciousness. Praeger.

Kunzendorf, R.G., Beltz, S.M. & Tymowicz, G. 1992. Self-awareness in autistic subjects and deeply hypnotized subjects: Dissociation of self-concept versus self-consciousness. *Imagination, Cognition and Personality* 11:129-41.

Munglani, R. & Jones, J.G. 1992. Sleep and general anesthesia as altered states of consciousness. *Journal of Psychopharmacology* 6:399-409.

Novak, P. 1996. Buddhist meditation and consciousness of time. *Journal of Consciousness Studies* 3:267-77.

Oakley, D. 1999. Hypnosis and consciousness: A structural model. *Contemporary Hypnosis* 16:215-223.

Pekala, R.J. & Kumar, V.K. 1989. Phenomenological patterns of consciousness during hypnosis: Relevance to cognition and individual differences. *Australian Journal of Clinical and Experimental Hypnosis* 17:1-20.

Pekala, R.J. & Cardena, E. 2000. Methodological issues in the study of altered states of consciousness and anomalous experiences. In (E. Cardena & S. Lynn, eds) *Varieties of Anomalous Experience: Examining the Scientific Evidence*. American Psychological Association.

Pekala, R. & Kumar, V. 2000. Individual differences in patterns of hypnotic experience across low and high hypnotically susceptible individuals. In (R. Kunzendorf & B. Wallace, eds) *Individual Differences in Conscious Experience*. John Benjamins.

Shapiro, D.H. 1982. Meditation as an altered state of consciousness: Contributions of Western behavioral science. *Journal of Transpersonal Psychology* 15:61-81.

Spivak, L., V. Puzenko, S. Medvedev, & Y. Polyakov 1990. Neurophysiological correlates of the altered state of consciousness during hypnosis. *Human Physiology* 16:405-410.

Tart, C.T. (ed) 1990. *Altered States of Consciousness* (third edition). Harper Collins.

Tart, C.T. 1998. Transpersonal psychology and methodologies for a comprehensive science of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Tart, C.T. 2000. Investigating altered states of consciousness on their own terms: State-specific sciences. In (M. Velmans, eds) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

Tinnin, L. 1990. Mental unity, altered states of consciousness, and dissociation. *Dissociation: Progress in the Dissociative Disorders* 3:154-59.

Travis, F. & Pearson, C. 2000. Pure consciousness: Distinct phenomenological and physiological correlates of "consciousness itself". *International Journal of Neuroscience* 100:77-89.

West, M. 1983. Meditation and self-awareness: Physiological and phenomenological approaches. In (G. Underwood, ed) *Aspects of Consciousness, Volume 3: Awareness and Self-Awareness*. Academic Press.

Wolman, B.B. & Ullman, U. 1986. *Handbook of States of Consciousness*. van Nostrand Reinhold.

Venkatesh S., Raju T.R., Shivani, Y., Tompkins G., & Meti B.L. 1997. A study of structure of phenomenology of consciousness in meditative and non-meditative states. *Indian Journal of Physiology and Pharmacology* 41:149-53.

Walsh, R. 1998. States and stages of consciousness: Current research and understanding. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

6.4f Parapsychology and Consciousness

Beloff, J. 1973. The subliminal and the extrasensory. *Parapsychology Review* 4:23-27.

Beloff, J. 1976. Mind-body interactionism in light of the parapsychological evidence. *Theoria to Theory* 10:125-37.

Beloff, J. 1980. Could there be a physical explanation for psi? *Journal of the Society for Psychical Research* 50:263-272

Beloff, J. 1987. Parapsychology and the mind-body problem. *Inquiry* 30:215-25.

Beloff, J. 1989. Dualism: A parapsychological perspective. In (J. Smythies & J. Beloff, eds) *The Case for Dualism*. Virginia University Press.

Bem, D.J. & Honorton, C. 1994. Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin* 115:4-18.

Bierman, D. 1998. Do psi phenomena suggest radical dualism? In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Blackmore, S. 1991. Psi in science. *Journal of the Society for Psychical Research* 57:404-11.

Blackmore, S. 1998. Why psi tells us nothing about consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Bohm, D.J. 1986. A new theory of the relationship of mind and matter. *Journal of the American Society for Psychical Research* 80:113-35.

Braud, W.G. 1994. The role of mind in the physical world: A psychologist's view. *European Journal of Parapsychology* 10:66-77.

Braude, S.E. 1979. *ESP and Psychokineses: A Philosophical Examination*. Temple University Press.

Braude, S.E. 1986. *The Limits of Influence: Psychokinesis and the Philosophy of Science*. Routledge and Kegan Paul.

Burns, J.E. 1993. Current hypotheses about the nature of the mind-brain relationship and their relationship to findings in parapsychology. In (K. Rao, ed) *Cultivating Consciousness*. Praeger.

Burns, J.E. 1993. Time, consciousness, and psi. In (B. Kane, J. Millay, & D.H. Brown, eds) *Silver Threads: 25 Years of Parapsychology Research*. Praeger.

Burns, J.E. 1986. Consciousness and psi. *PSI Research* 5:166-205.

Dilley, F.B. 1989. Mind-brain interaction and psi. *Southern Journal of Philosophy* 26:469-80.

Dilley, F.B. 1990. Telepathy and mind-brain dualism. *Journal of the Society for Psychical Research* 56:129-37.

Edge, H.L. 1989. Psi, self, and the new mentalism. In (L. Henkel & J. Palmer, eds) *Research in Parapsychology 1989*. Scarecrow Press.

Eisenbud, J. 1975. The mind-matter interface. *Journal of the American Society for Psychical Research* 69:115-26.

Goswami, A. 1986. The quantum theory of consciousness and psi. *PSI Research* 5:145-65.

Griffin, D.R. 1993. Parapsychology and philosophy: A Whiteheadian postmodern perspective. *Journal of the American Society for Psychical Research* 87:217-88.

Griffin, D.R. 1994. Dualism, materialism, idealism, and psi: A reply to John Palmer. *Journal of the American Society of Psychical Research* 88:23-39.

Grof, S. 2000. *Psychology of the Future: Lessons from Modern Consciousness*

Research_. State University of New York Press.

Heath, P. 2000. The PK zone: A phenomenological study. Journal of Parapsychology 64:53-72.

Honorton, C. 1985. Meta-analysis of psi ganzfeld research: A response to Hyman. Journal of Parapsychology 1:51-91.

Hubbard, T.L. 1996. Consciousness and cognition beyond the body: Functionalist cognitive science and the possibility of out-of-body experiences and reincarnation. Journal of the American Society for Psychical Research 90:202-20.

Hyman, R. 1985. The ganzfeld psi experiment: A critical appraisal. Journal of Parapsychology 49:3-49.

Hyman, Ray & Honorton, C. 1986. A joint communique: The psi ganzfeld controversy. Journal of Parapsychology 50:351-64.

Hyman, R. 1994. Anomaly or artifact? Comments on Bem and Honorto. Psychological Bulletin 115:19-24.

Jahn, R.G. & Dunne, B.J. 1987. _Margins of Reality: The Role of Consciousness in the Physical World_. Harcourt Brace Jovanovich.

Kreitler, H. & Kreitler, S. 1973. Subliminal perception and extrasensory perception. Journal of Parapsychology 37:163-88.

Krippner, S. & George, L. 1986. Psi phenomena as related to altered states of consciousness. In (B. Wolman & M. Ullman, eds) _Handbook of States of Consciousness_. van Nostrand Reinhold.

Mattuck, R. 1982. A crude model of the mind-matter interaction using Bohm-Bub hidden variables. Journal of the Society for Psychical Research 51:238-245.

Nash, C.B. 1976. Psi and the mind-body problem. Journal of the Society for Psychical Research 48:267-70.

Nash, C.B. 1995. A panpsychic theory of mind and matter. Journal of the Society for Psychical Research 60:171-73.

Poynton, J.C. 1994. Making sense of psi: Whitehead's multilevel ontology. Journal of the Society for Psychical Research 59:401-12.

Price, E.A. 1981. A "three worlds" perspective to the mind-brain relationship in parapsychology. Parapsychological Journal of South Africa 2:38-49.

Rao, K.R. & Palmer, J. 1987. The anomaly called psi: Recent research and criticism. Behavioral and Brain Sciences 10:539-51.

Rao, K.R. 1991. Consciousness research and psi. Journal of Parapsychology 55:1-43.

Rauscher, E.A. 1983. Multidimensional properties of consciousness and some laws of reality. PSI Research 2:53-66.

Richards, D.G. 1996. Psi and the spectrum of consciousness. Journal of the American Society for Psychical Research 90:251-67.

Roberts, F.S. 1991. Some apparently non-cerebral aspects of consciousness. Journal of the Society for Psychical Research 58:31-38.

Roberts, F.S. 1995. Is physically-based consciousness a reality? Journal of the Society for Psychical Research 60:398-400.

Roney-Dougal, S.M. 1986. Subliminal and psi perception: A review of the literature. *Journal of the Society for Psychical Research* 53:405-34.

Smythies, J.M. 1960. Three classical theories of mind. *Journal of the Society for Psychical Research* 40:385-397.

Stokes, D.M. 1982. On the relationship between mind and brain. *Parapsychology Review*, 13:22-27.

Stokes, D.M. 1993. Mind, matter, and death: Cognitive neuroscience and the problem of survival. *Journal of the American Society for Psychical Research* 87:41-84.

Stokes, D.M. 1997. *The Nature of Mind: Parapsychology and the Role of Consciousness in the Physical World*. McFarland and Co.

Tiller, W., Kohane, M., & Dibble, W. 2000. Can an aspect of consciousness be imprinted into an electronic device? *Integrative Physiological & Behavioral Science* 35:142-163.

Varvoglis, M. 1996. Nonlocality on a human scale: Psi and consciousness research. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness*. MIT Press.

Wade, J. 1998. Physically transcendent awareness: A comparison of the phenomenology of consciousness before birth and after death. *Journal of Near-Death Studies* 16:249-275.

Walker, E.H. 1984. A review of criticisms of the quantum-mechanical theory of psi phenomena. *Journal of Parapsychology* 48:277-32.

Woodworth, H. 1942. Report of investigations into an obscure function of the subconscious mind. *Journal of the American Society for Psychical Research* 36:185-230.

6.4g Phenomenology

Ackerman, D. 1990. *A Natural History of the Senses*. Randhom House,

Arvidson, P.S. 1992. On the origin of organization in consciousness. *Journal of the British Society of Phenomenology* 23:53-65.

Arvidson, P.S. 1996. Toward a phenomenology of attention. *Human Studies* 19:71-84.

Baars, B.J. 1993. Putting the focus on the fringe: Three empirical cases. *Journal of Consciousness Studies* 2:126-36.

Chokr, N.N. 1992. Mind, consciousness, and cognition: Phenomenology vs. cognitive science. *Husserl Studies* 9:179-97.

Deikman, A. 1996. 'I' = awareness. *Journal of Consciousness Studies* 3:350-56.

Depraz, N., Varela, F., & Vermersch, P. 2000. The gesture of awareness: An account of its structural dynamics. In (M. Velmans, ed) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

de Quincey, C. 2000. Intersubjectivity: Exploring consciousness from the second-person perspective. *Journal of Transpersonal Psychology* 32:135-155.

Diaz, J. 1996. The stream revisited: A process model of phenomenological consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a*

Science of Consciousness_. MIT Press.

Ellis, R. 1986. _An Ontology of Consciousness_. Kluwer.

Galin, D. 1994. The structure of awareness: Contemporary applications of William James' forgotten concept of "the fringe". Journal of Mind and Behavior 15:375-401.

Galin, D. 1996. The structure of subjective experience: Sharpen the concepts and terminology. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) _Toward a Science of Consciousness_. MIT Press.

Gallagher, S. 1997. Mutual enlightenment: Recent phenomenology in cognitive science. Journal of Consciousness Studies 4:195-214.

Gurwitsch, A. 1964. _The Field of Consciousness_. Duquesne University Press.

Gurwitsch, A. 1966. _Studies in Phenomenology and Psychology_. Northwestern University Press.

Ihde, D. 1977. _Experimental Phenomenology_. Putnam.

Jopling, D.A. 1996. Sub-phenomenology. Human Studies 19:153-73.

Koestenbaum, P. 1962. The sense of subjectivity. Review of Existential Psychology 2:47-65.

Lind, R. 1996. Micro-phenomenology: Toward a hypothetico-inductive science of experience. International Philosophical Quarterly 36:429-42.

Mangan, B. 1993. Taking phenomenology seriously: The "fringe" and its implication for cognitive research. Consciousness and Cognition 2:89-108.

Marbach, E. 1993. _Mental Representation and Consciousness: Toward a Phenomenological Theory of Representation and Reference_. Kluwer.

Marbach, E. 1996. Understanding the representational mind: A phenomenological perspective. Human Studies 19:137-52.

Marbach, E. 2000. The place for an Ego in current research. In (D. Zahavi, ed) _Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience_. John Benjamins.

Natsoulas, T. 1997. The presence of environmental objects to perceptual consciousness: An integrative, ecological and phenomenological approach. Journal of Mind & Behavior 18:371-390.

Natsoulas, T. 1997. The presence of environmental objects to perceptual consciousness: A difference it makes for psychological functioning. American Journal of Psychology 110:507-526.

Nelson, P. 1998. Consciousness as reflexive shadow: An operational psychophenomenological model. Imagination, Cognition and Personality 17:215-228.

Pekala, R.J., Wenger C.F., & Levine R.L. 1985. Individual differences in phenomenological experience: states of consciousness as a function of absorption. Journal of Personality and Social Psychology 48:125-32.

Pekala, R.J. & Levine, R.L. 1982. Mapping consciousness: Development of an empirical-phenomenological approach. Imagination, Cognition & Personality 1:29-47.

Rao, K.R. 1998. Two faces of consciousness: A look at Eastern and Western

perspectives. *Journal of Consciousness Studies* 5:309-27.

Shanon, B. 1984. The case for introspection. *Cognition and Brain Theory* 7:167-80.

Shear, J. 1996. The hard problem: Closing the empirical gap. *Journal of Consciousness Studies* 3:54-68. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press,

Stevens, R. 2000. Phenomenological approaches to the study of conscious awareness. In (M. Velmans, ed) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

Wilber, K. 2000. Waves, streams, states and self: Further considerations for an integral theory of consciousness. *Journal of Consciousness Studies* 7:145-176.

Varela, F. 1995. Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies* 3:330-49. Reprinted in (J. Shear, ed) *Explaining Consciousness: The Hard Problem*. MIT Press.

6.4h Foundations

Baars, B.J. 1994. A thoroughly empirical approach to consciousness. *Psyche* 1.

Baruss, I., & Moore, R.J. 1992. Measurement of beliefs about consciousness and reality. *Psychological Reports* 71:59-64.

Battista, J.R. 1978. The science of consciousness. In (K.S. Pope & J.L. Singer, eds) *The Stream of Consciousness: Scientific Investigation into the Flow of Experience*. Plenum.

Block, N. 2001. Paradox and cross purposes in recent work on consciousness. *Cognition* 79:197-219.

Conrad, D. 1996. Consciousness, privacy, and information. *Biosystems* 38:207-10.

Dennett, D. 2001. Are we explaining consciousness yet? *Cognition* 79:221-37.

Dunlop, K. 1912. The case against introspection. *Psychological Review* 19:404-13.

Flanagan, O.J. 1995. Consciousness and the natural method. *Neuropsychologia* 33:1103-15.

Foss, J. 2000. *Science and the Riddle of Consciousness: A Solution*. Kluwer Academic Publishers.

Goldman, A. 1997. Science, publicity, and consciousness. *Philosophy of Science* 64:525-45.

Goldman, A. 2000. Can science know when you're conscious? Epistemological foundations of consciousness research. *Journal Of Consciousness Studies* 7:3-22.

Grinker, R.R. 1953. Problems of consciousness: A review, an analysis, and a proposition. In (H. Abramson, ed) *Problems of Consciousness: Transactions of the Fourth Conference*. Josiah Macy Foundation.

Jack, A.I. & Shallice, T. 2001. Introspective physicalism as an approach to the science of consciousness. *Cognition* 79:161-196.

Lyons, W. 1986. *The Disappearance of Introspection*. MIT Press.

Miller, D. 2000. Designing a bridge for consciousness: Are criteria for a unification of approaches feasible? *Advances in Mind-Body Medicine* 16:82-89.

Nunez, R. 1997. Eating soup with chopsticks: Dogmas, difficulties, and alternatives in the study of conscious experience. *Journal of Consciousness Studies* 4:143-66.

Pekala, R. & Cardena, E. 2000. Methodological issues in the study of altered states of consciousness and anomalous experiences. In (E. Cardena & S. Lynn, eds) *Varieties of Anomalous Experience: Examining the Scientific Evidence*. American Psychological Association.

Scott, A.C. 1998. Reductionism revisited. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Scott, A. 2000. Modern science and the mind. In (M. Velmans, ed) *Investigating Phenomenal Consciousness: New Methodologies and Maps*. John Benjamins.

Searle, J. 1998. How to study consciousness scientifically. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Simon, H.A. 1997. Scientific approaches to the question of consciousness. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Stevens, S.S. 1966. Quantifying the sensory experience. In (P. Feyerabend & G. Maxwell, eds) *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*. University of Minnesota Press.

Varela, F. 1998. A science of consciousness as if experience mattered. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness 1996*. MIT Press.

Velmans, M. 1994. A reflexive science of consciousness. In *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

Velmans, M. 1996. Introduction to the science of consciousness. In (M. Velmans, ed) *The Science of Consciousness*. Routledge.

Velmans, M. 1998. Goodbye to reductionism: Complementary first and third-person approaches to consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Velmans, M. 2000. *Understanding Consciousness*. Routledge.

Wallace, B. 2000. *The Taboo of Subjectivity: Toward a New Science of Consciousness*. Oxford University Press.

Williams, D.C. 1934. Scientific method and the existence of consciousness. *Psychological Review* 41:461-79.

6.4i Consciousness and Science, Misc -----

Abramson, H.A. (ed) 1950. *Problems of Consciousness: Transactions of the First Conference*. Josiah Macy Foundation.

Abramson, H.A. (ed) 1951. *Problems of Consciousness: Transactions of the Second Conference*. Josiah Macy Foundation.

Abramson, H.A. (ed) 1952. *Problems of Consciousness: Transactions of the Third Conference*. Josiah Macy Foundation.

Abramson, H.A. (ed) 1953. *Problems of Consciousness: Transactions of the Fourth Conference*. Josiah Macy Foundation.

Abramson, H.A. (ed) 1954. *Problems of Consciousness: Transactions of the Fifth Conference*. Josiah Macy Foundation.

Bielecki, A., Kokoszka, A., & Holas, P. 2000. Dynamic systems theory approach to consciousness. *International Journal of Neuroscience* 104:29-47.

Blakemore, C. & Greenfield, S. 1987. *Mindwaves: Thoughts on Intelligence, Identity, and Consciousness*. Blackwell.

Bock, G.R. & Marsh, J. (eds) 1993. *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). Wiley.

Cohen, J.D. & Schooler, J.W. (eds) 1997. *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Cornwell, J. (ed) 1998. *Consciousness and Human Identity*. Oxford University Press.

Cotterill, R. 2000. *Enchanted Looms: Conscious Networks in Brains and Computers*. Cambridge University Press.

Hameroff, S.R, Kaszniak, A. & Scott, A. (eds) 1996. *Toward a Science of Consciousness: The First Tucson Discussions and Debates*. MIT Press.

Ito, M., Miyashita, Y., & Rolls, E.T. (eds) 1997. *Cognition, Computation, and Consciousness*. Oxford University Press.

Jarvilehto, T. 2000. The theory of the organism-environment system: The problem on mental activity and consciousness. *Integrative Physiological & Behavioral Science* 35:35-57.

John, E.R. 2001. A field theory of consciousness. *Consciousness and Cognition* 10:184-213.

Josephson, B. & Ramachandran, V.S. (eds) 1980. *Consciousness and the Physical World*. Pergamon Press.

Keyes, C. D. 1999. *Brain Mystery Light and Dark: The Rhythm and Harmony of Consciousness*. Routledge.

Marcel, A.J. & Bisiach, E. (eds) 1988. *Consciousness in Contemporary Science*. Oxford University Press.

Oakley, D.A. (ed) 1985. *Brain and Mind*. Methuen.

Scott, A. 1995. *Stairway to the Mind: The Controversial New Science of Consciousness*. Springer.

Sugarman, A.A. & Tarter, R.E. (eds) 1978. *Expanding Dimensions of Consciousness*. Springer.

Torey, Z. 1999. *The Crucible of Consciousness*. Oxford University Press.

Velmans, M. (eds) 1996. *The Science of Consciousness: Tutorial Essays*. Routledge.

Wilber, K. 2000. *Integral Psychology: Consciousness, Spirit, Psychology, Therapy*. Shambhala.

--

Compiled by David Chalmers, Department of Philosophy, University of Arizona.
(c) 2001 David J. Chalmers.

(Other) Online Bibliographies in Philosophy, Consciousness, and such

Compiled by [David Chalmers](#)

Philosophy of Mind

- [Color and Philosophy](#) (Alex Byrne and David Hilbert)
- [Commonsense Psychology](#) (Alexander Maeder)
- [Contemporary Philosophy of Mind: An Annotated Bibliography](#) (David Chalmers)
- [Internalism and Externalism](#) (Alberto Voltolini)
- [Knowledge Argument](#) (Torin Alter)
- [Language of Thought](#) (Larry Kaye)
- [Neurophilosophy](#) (Pete Mandik)
- [Nonconceptual Content](#) (Ron Chrisley)
- [Philosophy and AI](#) (Brian Rosmaita *et al*)
- [Philosophy of Psychiatry](#) (Christian Perring)
- [Self, Person, and Personal Identity](#) (Shaun Gallagher)
- [Self-Deception](#)

Other Philosophy

- [Bioethics](#) (LeRoy Walters and Tamar Joy Kahn)
- [Cognitive Science and Ethics](#) (Olle Blomberg)
- [Contextualism in Epistemology](#) (Keith DeRose)
- [Donald Davidson Bibliography](#) (Volker Meurer)
- [Environmental Ethics](#) (ISEE)
- [Evolutionary Epistemology](#) (Gary Cziko and Donald Campbell)
- [Metaethics](#) (James Lenman)
- [Objectivity](#) (Pete Mandik)
- [Online Papers on Charles Peirce](#) (Joseph Ransdell)
- [Philosophical Writing](#) (Will Buschert)
- [Philosophy of Artificial Life: Online papers](#) (Ezequiel Di Paolo)
- [Philosophy of Artificial Life](#) (Brian Keeley)
- [Philosophy of Chemistry](#) (HYLE)
- [Science Ethics](#) (Vince Hammer and Brian Tissue)
- [Top 100 Books in Analytic Philosophy](#) (Luciano Floridi)
- [Wilfred Sellars Online Papers and Bibliography](#) (Andrew Chrucky)

- [Women in Philosophy](#) (NOEMA)
- [Worlds and Modality](#) (Ted Sider)

Consciousness, Cognitive Science, etc.

- [ASSC Bibliography on Consciousness](#) (Thomas Metzinger and David Chalmers; password required)
 - [Artificial Life: Online papers](#) (Ezequiel Di Paolo)
 - [Autopoiesis and Enaction](#) (Randall Whitaker)
 - [Books on Mind and Related Topics](#) (Piero Scaruffi)
 - [Complexity Measures](#) (Bruce Edmonds)
 - [Consciousness and the Brain](#) (Ralph Ellis and Natika Newton)
 - [Evolutionary Psychology -- Online papers](#) (Ian Pitchford)(
 - [Online papers on Consciousness](#) (David Chalmers)
 - [Philosophical Issues in Connectionism \(Online Papers\)](#) (Joe Lau)
 - [Skeptic Bibliography](#) (CSICOP)
-

Go to:

- [David Chalmers' home page](#)

Philosophical humor

Compiled by [David Chalmers](#)

Some bits and pieces I've come across. Additions are always welcome

Lists and litanies

- [Proofs that P](#)
- [More proofs that P](#)
- [Philosophical kisses](#)
- [Why did the chicken cross the road?](#)
- [Non-philosopher's guide to philosophical terms](#)
- [Definitions from the profession](#)
- [Causes of death of philosophers](#)
- [Why no-one wants to play golf with a philosopher](#)
- [Philosophical warning labels](#)
- [Philosophical lexicon](#)

Jokes

- [A universal philosophical refutation](#)
- [Philosophy light-bulb jokes](#)
- [A Descartes joke](#)
- [Sensual guide to departments](#)
- [Philosopher's revenge](#)
- [Math/physics/philosophy joke](#)
- [Four philosophy jokes](#)
- [Profession jokes: Philosophers](#)

Satires

- [Brain in a vat at the wheel of a runaway trolley](#)
- [The Jean-Paul Sartre Cookbook](#)
- [A new refutation of the very possibility of Al Gore](#)
- [Husserl parody](#)
- [The Postmodernism Generator](#)
- [Realists Anonymous](#)

- [Jenny Jones on postmodernism](#)
- [Jerry Springer on philosophy](#)
- [The Lighter Report](#) (and the [Leiter report](#))
- [Wittgenstein on fog-like sensations](#)
- [Tech support, Nietzsche style](#)
- [On God's perfections](#), and [Is God made of soap?](#)
- [The beautification of logic](#)

Songs and poems

- [Philosophical Clerihews](#) (and [more](#))
- [Bruces' philosophers drinking song](#) (and [the Bruces sketch](#))
- [Philosophy songs](#)
- [I gave my love an emerose](#)
- [Richard Rogers' guide to nominalism](#)
- [Ballad of St. Anselm](#)
- [The Monads](#)

This and that

- [Analytic philosophy as reflected in the works of Monty Python](#)
- The philosophical truth about [cats](#), [dogs](#), [cockatoos](#), and [tiggers](#)
- [Everything is equally interesting](#)
- [Horoscopes for philosophy graduate students](#)
- [Socrates' Argument Clinic](#)
- [Learn logic with Beavis and Butthead](#)
- [I also dated Zarathustra](#)
- [Self-referential story](#)
- [They're made of meat!](#)
- [Goat philosophy](#)
- [New logical particles](#)
- [Tractatus Logico-Randomus](#)
- [Para-frays, or, The writes of passage](#)
- [Zombies on the web](#)
- [Introductory philosophy quiz](#)
- [Philosophy Comics](#)

See also

- [Philosopher Jokes](#) (Miami)
- [Philosophy Humor](#) (Valdosta)
- [Chomskybot](#)
- [Economist Jokes](#)
- [Mathematical Jokes](#)
- [Mathematical Humor](#)
- [Mathematical Big-Game Hunting](#) (see also [Elephant Hunting](#))
- [Find your Birthday in Pi](#)
- [Physics Humor](#)
- [Physics Humor Page](#)
- [Psychology Humor Page](#)
- [Science Humor](#)
- [Twinkies Project](#)
- [Viola Jokes](#) & other [Music Jokes](#)

Go to:

- [David Chalmers' home page](#)

Proofs that p

Davidson's proof that p :

Let us make the following bold conjecture: p

Wallace's proof that p :

Davidson has made the following bold conjecture: p

Grunbaum:

As I have asserted again and again in previous publications, p .

Putnam:

Some philosophers have argued that not- p , on the grounds that q . It would be an interesting exercise to count all the fallacies in this "argument". (It's really awful, isn't it?) Therefore p .

Rawls:

It would be nice to have a deductive argument that p from self-evident premises. Unfortunately I am unable to provide one. So I will have to rest content with the following intuitive considerations in its support: p .

Unger:

Suppose it were the case that not- p . It would follow from this that someone knows that q . But on my view, no one knows anything whatsoever. Therefore p . (Unger believes that the louder you say this argument, the more persuasive it becomes).

Katz:

I have seventeen arguments for the claim that p , and I know of only four for the claim that not- p . Therefore p .

Lewis:

Most people find the claim that not- p completely obvious and when I assert p they give me an incredulous stare. But the fact that they find not- p obvious is no argument that it is true; and I do not know how to refute an incredulous stare. Therefore, p .

Fodor:

My argument for p is based on three premises:

1. q
 2. r
- and

3. p

From these, the claim that p deductively follows. Some people may find the third premise controversial, but it is clear that if we replaced that premise by any other reasonable premise, the argument would go through just as well.

Sellars' proof that p :

Unfortunately limitations of space prevent it from being included here, but important parts of the proof can be found in each of the articles in the attached bibliography.

Earman:

There are solutions to the field equations of general relativity in which space-time has the structure of a four- dimensional Klein bottle and in which there is no matter. In each such space-time, the claim that not- p is false. Therefore p .

Goodman:

Zabludowski has insinuated that my thesis that p is false, on the basis of alleged counterexamples. But these so- called "counterexamples" depend on construing my thesis that p in a way that it was obviously not intended -- for I intended my thesis to have no counterexamples. Therefore p .

.

Outline Of A Proof That P [\(1\)](#): Saul Kripke

Some philosophers have argued that not- p . But none of them seems to me to have made a convincing argument against the intuitive view that this is not the case. Therefore, p .

(1) This outline was prepared hastily -- at the editor's insistence -- from a taped manuscript of a lecture. Since I was not even given the opportunity to revise the first draft before publication, I cannot be held responsible for any lacunae in the (published version of the) argument, or for any fallacious or garbled inferences resulting from faulty preparation of the typescript. Also, the argument now seems to me to have problems which I did not know when I wrote it, but which I can't discuss here, and which are completely unrelated to any criticisms that have appeared in the literature (or that I have seen in manuscript); all such criticisms misconstrue my argument. It will be noted that the present version of the argument seems to presuppose the (intuitionistically unacceptable) law of double negation.

But the argument can easily be reformulated in a way that avoids employing such an inference rule. I hope to expand on these matters further in a separate monograph.

Routley and Meyer:

If $(q \ \& \ \text{not-}q)$ is true, then there is a model for p . Therefore p .

Plantinga:

It is a model theorem that $p \rightarrow p$. Surely its possible that p must be true. Thus p . But it is a model theorem that $p \rightarrow p$. Therefore p .

Chisholm:

P -ness is self-presenting. Therefore, p .

Morganbesser:

If not p , what? q maybe?

A Universal Philosophical Refutation

A philosopher once had the following dream.

First Aristotle appeared, and the philosopher said to him, "Could you give me a fifteen-minute capsule sketch of your entire philosophy?" To the philosopher's surprise, Aristotle gave him an excellent exposition in which he compressed an enormous amount of material into a mere fifteen minutes. But then the philosopher raised a certain objection which Aristotle couldn't answer. Confounded, Aristotle disappeared.

Then Plato appeared. The same thing happened again, and the philosophers' objection to Plato was the same as his objection to Aristotle. Plato also couldn't answer it and disappeared.

Then all the famous philosophers of history appeared one-by-one and our philosopher refuted every one with the same objection.

After the last philosopher vanished, our philosopher said to himself, "I know I'm asleep and dreaming all this. Yet I've found a universal refutation for all philosophical systems! Tomorrow when I wake up, I will probably have forgotten it, and the world will really miss something!" With an iron effort, the philosopher forced himself to wake up, rush over to his desk, and write down his universal refutation. Then he jumped back into bed with a sigh of relief.

The next morning when he awoke, he went over to the desk to see what he had written. It was, "That's what *you* say."

[From Raymond Smullyan, *5000 B.C. and Other Philosophical Fantasies*. St. Martin's Press, 1983]

Math, Physics, & Philosophy

Dean, to the physics department. "Why do I always have to give you guys so much money, for laboratories and expensive equipment and stuff. Why couldn't you be like the math department - all they need is money for pencils, paper and waste-paper baskets. Or even better, like the philosophy department. All they need are pencils and paper."

Philosophical Clerihews

By Dean Zimmerman

I've wasted an inordinate amount of time on the lowest verse form in existence, the Clerihew. Here's the cream or the dregs, depending on how you look at it.

The first one ever written (1890, by the 16-year-old E. Clerihew Bentley) went like this:

Sir Humphry Davy
Detested gravy.
He lived in the odium
Of having discovered Sodium.

Here's a really good one by Maurice Hare:

Alfred de Musset
Used to call his cat Pusset.
His accent was affected.
That was to be expected.

The *Poet's Manual and Rhyming Dictionary* of Frances Stillman defines the clerihew as "a humorous pseudo-biographical quatrain, rhymed as two couplets, with lines of uneven length more or less in the rhythm of prose. It is short and pithy, and often contains or implies a moral reflection of some kind. The name of the individual who is the subject of the quatrain usually supplies the first line."

Bentley himself wrote a few on philosophers. W. H. Auden has a book of clerihews with ones on Kierkegaard, Hegel, and a couple of others. Paul Horgan's book of clerihews has a few as well. By and large the ones on philosophers aren't very good. The best I know of are these:

Desiderius Erasmus
Suffered from one of the rare asthmas.
His worst wheezes
Were caused by over-ripe cheeses.

(Paul Horgan)

John Stuart Mill,
By a mighty effort of will,
Overcame his natural bonhomie

And wrote *Principles of Political Economy**

*-`With some of their Applications to Social Philosophy'.

It was a weakness of Voltaire's
To forget to say his prayers,
And one which to his shame
He never overcame.

Sir James Jeans
Always says what he means;
He is really perfectly serious
About the Universe being Mysterious.

(all by Bentley)

But the largest batch of philosophical clerihews follows...

Philosophical Clerihews

by Dean W. Zimmerman

This form was evidently
invented by E. Clerihew Bentley.
He did nothing else well,
but what the hell?

Moses Maimonides
wrote vast quantities
and stood for amity
in an age of Kalamity.

G. E. Moore
was a bit of a bore;
more an old fossil

than a Cambridge Apostle.

By ghosts, C.D. Broad
was not generally overawed;
but he looked a trifle haggard
after his Examination of McTaggart.

Fearing Wittgenstein was blotto,
Popper adopted the motto:
"I will not argue with a joker
who's brandishing a red-hot poker."

"Peter Geach!" said God,
"would it not be slipshod
were I to raise not the authentical
you, but someone relatively identical?"

Roderick Chisholm
was never accused of asceticism
by Roderick Firth,
who was equally down-to-earth.

Nicholas Rescher
is under no pressure
to write another book;
but look!

Edwin B. Allaire
prefers particulars bare;
he doesn't find them rude

in the nude.

Jaegwon Kim,
writing poetry under a psuedonym,
met with considerable inconvenience
in rhyming "supervenience".

"I sneezed", explained Al Plantinga,
"as I scaled the rockface, barely panting. 'Ge-
sundheit!' said a voice; but not God's, inasmuch
as that's German not Dutch."

If your battery's dead, don't ask Bradley
to give you a jump. He'll say: "Gladly,
I have jumper cables; but need another item:
a cable to connect the cables" and so ad infinitum.

The British railroad
was the ruin of C.E.M. Joad.
Back then it wasn't cricket
to be caught without a ticket.

Jose Benardete
cooks large batches of spaghetti.
His noodles make you full
when you've eaten Aleph-null.

Alfred Freddoso
thinks his own work's just so-so;
he'd never go so far as

to compare it to Suarez's.

Sydney's David Malet
Armstrong had little palate
for Marxism;
and that'll spark schism.

Derek Parfit
wouldn't wear a scarf. It
made him more bold
to think someone else would catch cold.

G. H. von Wright
is not very strict.
He remains polite
when called "von Wright".

Here's a moral pilfered
from Wilfrid
Sellars: Have enough to drink
and all ice cubes look pink.

Wilfrid Sellars
abhorred poor spellers;
they made his blood congeal
and gave him a raw feel.

When he was younger,
the ignorance of Peter Unger
was vast;

but it didn't last.

A handy acronym
invented by Colin McGinn
explains why
philosophers shouldn't even try.

There's no disputin'
that Grigori Rasputin
had more will to power
than Schopenhauer.

Susan K. Langer
liked instruments that twang. Her
Philosophy in a New Key
was dedicated to Paul Stookey.

Unfortunately, Thales
never tried Bailey's.
It's smoother than water, so better suited
to be that of which all is constituted.

The metaphysics of Ernie Sosa
has little in common with that of Spinoza,
except for this: both claim to show
all things are like a ball of snow.

Although it hurt Curt Ducasse
to be kicked in the ass,
he was filled with elation

at the observability of the causal relation.

Escaping at night from the embalmer's,
The zombies sought help from Dave Chalmers.
Though their speech was mere echolalia,
He knew what they wanted: dancing qualia.

With parenthetical suggestions for cartoons:

(Gritting his teeth, fingers twitching)

Stephen Stich
wouldn't scratch an itch:
one mustn't accept the ontology
of common sense psychology.

(Picture of a bat with Nagel's face, putting wing over face in vampire fashion to avoid looking at a bust of Hegel)

Don't ask Thomas Nagel
to study Hegel;
he'd rather be a bat
than do that.

(Dancing waltzily while sensing redly)

Michael Tye,
while just a young guy,
had an enthusiasm for the adverbial
that was proverbial.

(In Saturday Night Fever outfit at a disco, opening letter and looking depressed)

In youth, Frege
could really shake a leg. A
correspondence with Russell
left him unable to do The Hustle.

And here are some by Brian Leftow:

Who better to get my first clerihew
Than Yale's new chair, Robert Merrihew
Adams? His job is one to admire. He
Should hire me.

Goodman trades classes
for all possible scattered masses:
No matter where the parts are from,
they are many, ergo sum.

Hobbes has no abstracta
Just entia tacta-
what isn't in chunks
flunks.

"To be or not to be?"
Said Parfit, "totally
redundant. Hamlet pushed the issue way too far.
What matters isn't B, but R."

That Alvin
Rhymes with Calvin
Is, in my estimation,
Proof of predestination.

Quine and Cartwright point with glee
to the 1992 *Times*, where we
read that Mr. Bobbitt
was for awhile a scattered object.

Richard Rorty
Wrote philosophy till forty.
Then midlife crisis hit
And he quit to write... pragmatism.

If reference is inscrutable
And ostension mutable
There's just no knowing
where Quine is going.

When reading David Lewis
One can't say what the thing to do is-
Gape? Guffaw? Collapse and fall?
Elsewhere some counterpart does them all.

G.W.F. Hegel
Is less nutritious than a bagel.
A bagel has just one hole in the dough-
Hegel's all holes, wherever you go.

Antony Flew
Wouldn't be caught dead in a pew.
A wiser head
Would worry that some day he'll be caught dead.

According to W.V. Quine
Any ontology's fine
And that's why I
Think he's a heckgavagai.

According to John Stuart Mill,
Better check whether 2+2 still
Make four. They did today,
But as for tomorrow - who can say?

Nietzsche thought that all events
Repeat themselves- which makes no sense.
Nietzsche thought that all events
Repeat themselves- which makes no sense.

One more by Tom Kirby-Smith:

Ludwig Wittgenstein
Hardly ever went out to dine.
Be the menu never so abundant,
He found "green leafy lettuce salad" tautological and redundant.

I Also Dated Zarathustra

[Sharon Wahl](#)

On the Dating Game

It's not the questions that count, and it's not exactly the answers. You ask a silly thing: "What kind of fruit would you be?" and you hope a voice on the other side of the dividing panel, one from the row of three men on stools the audience can see but you cannot, will reach through the make-up and studio laughter to give a sign that he knows, yes, this is stupid, but what we want is not stupid: who, after all, knows how to find the person he will love?

I had looked nearly everywhere else and decided that if necessary, and it seemed to be necessary, I would look here, too. I would sit in a tight short dress with my legs placed at an advantageous angle to the camera, crossed at the knees with one high heel dangling from my toes. It's a favorite pose and successful. The men I was quizzing couldn't see it but the cameraman was driven crazy. With each question I asked, he returned to dwell on the ankle and bare heel.

Well then. What kind of fruit *would* he be?

Bachelor Number One? "Something that will make a really great pie. And that's a promise. I am one dependable fruit."

Bachelor Number Two? "I'd be plums, I'd be sweet and red and *very* juicy. And honey, I'm always ripe."

Bachelor Number Three? "I am a north wind to ripe figs. I am a prophet of the lightning and a heavy drop from the clouds. I am an intoxicated sweet lyre-- a midnight lyre, a croaking bell which no one understands but which still must speak!"

I didn't move but I was suddenly aware of my thighs, of the insides of my arms. Bachelor Number Three had a voice like a cloud speaking, traces of roar and thunder and waves held together with honey-cello. But what did I know about him? He might be ridiculous, I thought. He might be sublime.

When a man is mysterious enough, when I have no idea which things will be good or bad or where the problems will be or even what will happen next, it makes me think that anything might be possible.

"Num Ber Two! Num Ber Two!" the female audience chanted during the Thinking Music.

Number One hugged and kissed me. Number Two slipped a note with the number of his hotel room down the front of my dress. Number Three stiffly bowed.

He was older than the others. His eyebrows were shaggy and his mustache needed trimming. He wore a long black coat, which would never, in the course of our acquaintance, be removed.

He bowed and shook my hand. Then, lifting his eyebrows, he peered at it: a hand, yes, in his hand, a hand attached to other parts of a fairly beautiful woman with whom he was about to spend a weekend in Las Vegas. That is where they were sending us, to the glittering neon desert, to try our luck.

"Ah," he said. "What a long and beautiful hand! It is the hand of one who has always distributed blessings. But now it holds fast him you seek, me, Zarathustra."

On Love of the Strange, and of Men

Oh, it's a love of aching things, not sweet things, a love of stars, long silences, birds that sing at night. A love like an old Victorian house that is almost empty: a winding staircase to a room paneled in dark wood, its worn polish dimly reflecting light from a chandelier with yellow bulbs, a few notes of Chopin played in another wing of the house by an overnight guest who does not speak. (I would love him, too, thin like a carved rail of the staircase, awkwardly holding a glass of wine at dinner so that you might think his hands hadn't comfortably mastered the world; and yet, the Chopin.) A love like stamps from a country whose name you don't recognize, exquisite writing in a script you can't read or a language you don't know.

I have never loved anything I've understood.

"All mousetraps of the heart have now again been set! And wherever I raise a curtain, a little night-moth comes fluttering out." I was on the balcony, standing in the hot sun looking out over the Nevada desert, where the bare mountains were made beautiful by shadows. Zarathustra, in the hotel room, stretched his hands to the blasting air conditioner and spoke to himself.

The strange are often the best defended; that is how they have kept themselves strange. "I am not on my guard against deceivers, I must be without caution," Zarathustra said. Their hearts are not guarded deliberately, but by being essentially impenetrable.

It's a love that is like loneliness.

Of Balancing on Four Legs

It was a warm night, with lots of neon. It was one of those nights when the world seemed to be made not of people, but of couples. Everywhere I looked bodies were paired together, connected at the hands or more tightly around the waist, awkward animals walking with a tilt and lean, off-balance, unsymmetrical. They reminded me of those children's books with the pages cut in three, each section the top, middle, or bottom of an animal, so that the normal old heads and legs and bellies could be made into sillier

creatures: a salamousowl, a girelephish, a pandazebrogator.

And yes, I wanted to be part of it all. Of course I did. Oh, to be damply interlaced at the palms. To be affectionately leashed, tethered in the crowded streets, appended. To make a wider obstacle on the sidewalk, a wandering self-absorption that others had to navigate, rather than this narrow thing that darts and slips politely by.

But Zarathustra understood nothing of this. He would walk between or duck under the arms of people clearly together, something I by instinct could not do. It was impossible to truly accompany him, to predict and accommodate his walking speeds or stopping places. He seemed to resent being tied down, even by gravity. He walked with high fast steps and frequently bumped into things. Really he couldn't see very well; until he was at arm's length he wasn't completely certain what he was looking at. This meant that everything out of reach was immensely interesting, and the things close by merely obstacles.

And so we wandered, over the Brooklyn Bridge, into the great pyramids. When he pushed people aside to get a better view of the volcano or the pirates, I apologized and petted them, winked and shrugged: he's hopeless, what can I do? And really most people didn't mind so much once they saw him. He was so completely taken over by the spectacle, so silly looking and so utterly happy, that often enough they ended up watching us instead.

If there were a Couple Game, a prize for the oddest match on the Vegas strip, we would have won. Zarathustra could have been my father. Or crazy. Perhaps he was extremely wealthy? Something the casinos had flown in. And surely I was being paid. Madmen didn't get girls with legs this long. The casino had hired me, to raise the stakes. Or he had hired me, for much the same function. Oh wasn't he a lucky man!

We just didn't look like we were on a date. I wondered if this was why the show had assigned us a chaperone-- to give Zarathustra a backup option, and to let me off the hook. Surely this odd man couldn't be my type?

"Perhaps you would like a game of poker?" the chaperone had suggested, after dinner.

"I have played dice with the gods at their table, the earth," said Zarathustra.

"Oh, really? And where was this?"

The chaperone was a motherly type. She was a mother-in-law on a honeymoon who wanted desperately to be a fairy godmother. We had tickets to a show, coupons for free drinks, a hundred dollars in betting tokens, all of which she had bestowed upon Zarathustra, all of which had entered the pockets of his coat, never to be seen again.

She had tried to take the coat from him back in the hotel room. It was thick black wool, buttoned from

his chest to his knees, and must have felt like wearing a portable oven. Still, it seemed possible that he had simply forgotten it was there. The chaperone suggested, coquettishly reaching for the top button, that she would make him more comfortable. Zarathustra said to her, "This is the tarantula's cave! Do you want to see the tarantula itself?" She didn't. She never mentioned it again.

I had said nothing. By then I'd come to enjoy the prospect of removing the coat myself, at the end of the night. I pictured my long arms snaking from behind, my bare skin on his black wool, undoing him button by button.

And what was Zarathustra thinking, when he looked at me? Did he like tall redheads? Did he wish I was wearing more? "You have the most lovely shoulders," as the chaperone put it, winking at Zarathustra, who missed his cue for a courtly compliment.

It was difficult to imagine either of them on a date. She needed to have always been married; he needed to be taken by the arm and gently steered, directed this way or that with no mention made of the change of pilot. Or so I had thought; that was an innocent time, early evening.

"Try your luck!" the chaperone said, giving each of us a quarter as we passed a row of slot machines. They jingled and blinked happily among themselves like well-fed babies in need of occasional burping.

Zarathustra fed the nearest one his quarter, and I dropped mine in after, doubling the bet. The dizzy smile started up, spinning to lock on cherry; cherry; cherry. Quarters poured out. It was a good sign, I thought.

Zarathustra picked up a quarter and gave it back to the chaperone, then walked on. The chaperone had a wordless fit: sputter! etc. She saw that yes, we really did intend to walk away. This was her chance. A thousand quarters: she looked hungrily at the pile, thinking that among so many there must be others that would win, she had only to find them, to kiss each frog and find the princes.

And so we left her. We didn't need a chaperone; we had his coat.

We walked for hours. The signs, the palm trees, the cars, the casinos, the people he passed on the street and the things they carried, all seemed to Zarathustra equally there for entertainment. The boy whose cotton candy he sampled (gazing in rapture at the pink cloud, plucking a wisp from it, licking tentatively at the barely there) had no right to scream at him, no, not when one possessed an object so delicious! Not when one possessed an object he wanted. I wasn't able to explain this to the boy or his father, but bought the boy another cone and a small teddy bear, hoping he wouldn't notice how much it resembled Zarathustra.

When I caught up again a uniformed chauffeur had handed Zarathustra a vividly illustrated brochure of a desert brothel, and was offering to drive him there. Zarathustra seemed confused that the pamphleteer was not, as he had naturally (so he insisted) assumed, a philosopher or revolutionary. Clearly anyone distributing printed material deserved his full attention.

"Ah," he said, standing in a bright spotlight to study the pamphlet. As he considered the transcendent ideologies of Daisy and Trixie, his shadow roamed over the side of a building across the street. He raised his arms and the shadow tried to climb through the window of a hotel room much too small for it. Two blonde teenaged girls approached and the arms went after them, straining for a squeeze, wiggling fingertips poked their knees and tried to trip them, the pamphlet waved madly over their bodies as they escaped unclutched. The shadow shrugged. Zarathustra shook his head at it disapprovingly. "I must keep it under stricter control-- otherwise it will ruin my reputation."

I pulled him away by the arm, and found a moth hole in the outside of his coatsleeve, which fit my little finger like a ring. Z didn't seem to notice, so I kept it there. By this time it was becoming quite clear how much he needed me.

Of Dancing Things

"Listen," said Zarathustra. "It is night: now do all leaping fountains speak louder." We were sitting side by side on the low concrete ledge of the fountains in front of our hotel.

"And what do they say?"

"They are deep but without thoughts, like little secrets, like after-dinner nuts."

Zarathustra had his hand in the shallow water, plucking it with his finger, a soft plink plink, plink plink.

I slipped off my shoes, and stepped in. My legs were a lovely blue in the reflected light. The pool was ankle deep, gently swirled and bubbled by the water spraying and tumbling in the middle.

Zarathustra scooped some pennies from the bottom of the pool, a few of the good luck pennies which were everywhere, and dropped them in again. They plopped straight back to the bottom, shy as frogs.

"I know what the water says. It says it's given up waiting for the pennies to swim, it wants people to throw goldfish, instead."

"Everything is asleep," Zarathustra said. "Even the water is asleep. Its eye looks at me drowsily and strangely. But it breathes; I feel it. And it is dreaming, look how it tosses and turns."

The air was warm, the water was warm. My feet were delighted. I walked through the shallow pool to the other side of the fountains on my blue legs, nudging pennies with my toes. Which felt better, stork steps, or fish steps? Which felt better, going from the air to the water, or the water to the air?

Zarathustra couldn't see it, but I was dancing. The water's hands slipped teasing through my fingers. My tall partners the palms swayed out-of-reach above me. The desert breezes came in now and then,

stretched across us like a chorus line. Each of us with our private shivers.

I thought it would be nice, on a night like this, to fly. So I left the water in a balletic leap, toes neatly pointed, one arm up, one to the side. But when I came down I slipped on the pennies. I slid into a sharp pain-- a bottlecap. My foot was cut. Zarathustra, in his brown leather boots and thick coat, came striding through the water to see what I was swearing at.

The cut wasn't deep; a small bleeding gouge on the side of my big toe. I dried it with tissue, and found a band-aid in my purse. But Zarathustra was at war: there were to be no bottlecaps among the pennies. He went down on hands and knees to search them out. His coat settled into the water reluctantly, floating and full of pocketed air, then sponging and swelling and sinking.

At first he was amusing, almost gallant. The eradication of sharp edges, the world made safer for toes. But. He kept on and on, crawling through the water, attacking nickels then releasing them, puzzling over gum wrappers and bits of palm leaf, banishing all manner of suspicious things to his wet pockets.

I wanted to go. I wanted a drink, and I said so. I wanted a cold, numbing, double margarita. Zarathustra ignored me, or maybe didn't hear. I said this louder. Putting on my shoes made my toe hurt more. "Come on," I said. Zarathustra tucked another bottlecap into his coat pocket. Probably they were falling back out again as he crawled. "This is ridiculous!" I shouted.

"All fish talk like that; what they cannot fathom is unfathomable," he said.

"I'm going inside," I said. At the hotel entrance I turned back to see if he had followed. He hadn't. He lay in the fountain stretched full length, legs spread, arms wide, his hands flapping and spanking the water, cooling his overheated addled brain and burbling.

On the Compassion of the White Tigers

Sometimes it helps to be drunk. One stiff margarita and I remembered how much I liked being wet, how charming I would look with my hair thoroughly soaked and shaken into ringlets. The fountain would tell me so itself, really, I was ready to hear all sorts of things. I was ready to sing along.

But Zarathustra had finished his bath. I found him wandering the main corridor of the hotel, waving his arms and spraying drops like a wet, happy dog.

I preceeded him at a safe custodial distance. He didn't notice me; being human was sufficient camouflage.

The corridor led past a large white room, the mountain-kingdom throne room of the resort's white tigers, and there Zarathustra came to a sudden, astounded stop. Two white tigers lay on the floor not far from him, gazing into a distance we were certainly no part of.

"Ah, my brothers! If only my lioness Wisdom had learned to roar fondly!"

He sat down on the floor of the marble walkway and spoke to them, earnestly, and at great length.

Their heads went slowly side to side, looking at everything as though they already knew it in their painted kingdom of empty white spaces and artificial things, snow in their minds. They both had beautiful intelligent expressions and seemed complete, resigned to a stable satisfaction, not requiring further enlightenment.

He told them everything. I leaned half-behind a marble pillar and listened.

"I have always wanted to caress every monster. A touch of warm breath, a little soft fur on its paw-- and at once I have been ready to love and entice it. Love is the danger for the most solitary man, love of any thing *if only it is alive!*"

With my eyes closed, without the dreadful evidence of the wet coat, without the spreading streams of water the other casino-goers were gingerly stepping around as they passed him, looking at him and also trying not to look, not to spoil the night's glitter, not to wet their shoes, but especially not to hear him, surely this was gibberish, a language they didn't share, an animal speaking to animals, they pretended he was grunting--

With my eyes closed there were only his words, his words like little spotted night moths, and his low voice. The most enticing sort of monster is a soul aching with ferocious tenderness.

"To be sure, I am a forest and a night of dark trees: but he who is not afraid of my darkness will find rosebowers too under my cypresses."

Did I think I could change him? Yes. Did I care what the others thought, the tip-toers, the avoiders of puddles? No. Did I think he would escort me through the dark cypresses into his rose bower? Yes. I pictured pale yellow climbing roses, a weathered bench strewn with fallen petals. I thought he had made the bower with its little bench so that I would have a place to sit.

"Ah, my friends, it is the evening that questions thus within me. Forgive me my sadness! We were made for one another, you gentle, strange marvels. And we have already learned so much with one another!"

He left them, much cheered and quieted, though still dripping, and followed me into the nearest bar.

On Learning to Drink from All Glasses

Zarathustra was fascinated by the little white and red striped straw that came with his margarita. It made

slurps in such a nice assortment, hollow air sucks, loud bubble-burble, liquid squishes. It suffered lime-pulp obstructions, it poked under cubes of ice, it vacuumed the last stray drops. He kept the straw in place between his lips and sipped liked a bird, lowering his head to the glass, tilting side to side, poking and hunting.

Finally the ice was sucked tasteless. But Zarathustra hadn't had nearly enough of the straw. Here and there on the empty tables were glasses left behind, some with enough color to be more than ice melt. Come to us, they called, we are diluting. The straw perked like a lower appendage. He couldn't resist. "He who does not want to die of thirst among men must learn to drink from all glasses," he said as he left me.

But most of these abandoned drinks were only drowned remains, melted ice and cherry stems and gnawed wedges of orange or pineapple. They were insufficiently alcoholic. He saw a fresh cocktail with a pink umbrella on another table, and with his straw, went directly to it. The woman who had ordered the drink was horrified. She stared at the glass as though a man had popped out of it, not into it. It might happen again, acrobats could leap from the glass and stick their daiquiri-flavored tongues into her open mouth.

The couple at the next table said that someone really ought to call security. So I told them, leaning over confidentially, that Zarathustra was a paid comedian. "Yes, it's part of his act," I said. And truly, this was a stroke of genius.

"Oh!" they said, of course, it all made sense, and then they excitedly told the next couple. The world is so much nicer when it's making sense. I could see the rumor navigate the bar, who had heard and who hadn't, by the shift from flinches to smiles, then even to competition for Zarathustra's attention. Men called him over to shock their wives, a giggly, safely titillating little shock, and women offered up their drinks to him as he passed.

Zarathustra was delighted. Here was the world as he had intended it to be, here he was understood! Drunk, he was the perfect entertainer, a cross between a host and a clown, offering toasts and renaming all the drinks. The Death of God! he pronounced. The Will to Power! Live Dangerously! Eternal Recurrence!

The names were a bit mystifying to the drinking folk, but they loved him. He leaned over to suck another drink and lost his balance in a plump, elderly lap. The plump woman laughed and tried to grab him. "O Earth, you have grown too round for me!" he cursed. "Live Dangerously!" they toasted. And to me, "He's good, isn't he? It's very clever."

Finally he came back to our table. He was merry and elated and fidgety, reaching out when fresh drinks passed, glazing over with alcohol and exhaustion. It was getting late. I was ready for a little gratitude. I moved my chair around the table very close to his. He still had the perky little straw in his mouth.

"You haven't yet tasted me," I said, and sucked on the end of his straw. It pinched his tongue, and he spat it out.

"Oh, your straw!" I said. It had fallen to the floor on the other side of his chair. I leaned over to get it, full across his lap, and while I was there I took the time to undo the two lower buttons of his cool damp wool cocoon. He was absolutely still. Inside and out. But he thanked me for the straw.

"Zarathustra," I said, sort of snuggling up to him but distracting him by pointing to a young couple kissing vigorously in a dark corner of the bar. "Look. What are they doing?"

"They are discovering new words."

"You mean, speaking with different tongues?"

"Before long they will be devising festivals!"

"I like it when they smile between kisses."

"Like cats they arch their backs, they purr."

"Well, look where his paw is. . ."

"You must not want to see everything."

"But I do. I think I do want to see everything, don't you?"

"For that you must have long legs."

"Let's see." I hooked my foot under his and stretched our legs out, to measure them side by side. "Mine is longer! I thought so."

"You are making this cave sultry and poisonous, you evil sorcerer!"

"Sorcerer!"

"If I may tickle you with this name."

"I would prefer fingers."

Someone had left a marble in the ashtray on our table, a milky blue color, which Zarathustra began to roll across the table. I said it was the moon's right eye, sent to spy on us.

"He is lustful and jealous, the monk in the moon," Zarathustra said.

"And what about you?"

"I am invulnerable only in my heels."

"Are you lustful? Do you like the girl over there, kissing?"

"One should speak about women only to men."

"Why? Tell me! What do you think of women?"

"They know how to blow horns and to go around at night and awaken old things that have long been asleep."

"Zarathustra," I said, "what are you hiding under your coat?"

"It is a little truth that I carry. But it is as unruly as a little child, and if I do not stop its mouth it will cry too loudly."

"Ah," I said. "Oh."

"You want to call it by a name and caress it. You want to pull its ears and amuse yourself with it."

"Yes," I said, "you're right, I do." I suggested that it was time to go back to the room.

Of an Introduction to the Ideal Woman

I led our trek from bar to bedroom through a small dense jungle, under waterfalls, over a green and gold wooden elephant, through beaded curtains into a dim red room painted with a harem scene. We leaned against the wall of bare-breasted dark-eyed women and pretended to choose a favorite, whispering to each other in the opium light, bumping and brushing hands in the purple shadows.

And then I found the perfect finale. On a pedestal advertising the week's performers was a classic chorus girl mannequin, leg kicked high, arms wide. What a fetching pose, I thought. I could do that, too. A private Vegas floor show for Zarathustra.

I climbed up and danced a can-can with her, kicking my equally long legs, my arm across her shoulders. She was so cute. She was terribly sweet and happy. And Zarathustra gave us just the look I wanted, he gazed adoringly, ecstatically at the two of us intertwined. He said exactly what I wanted him to say:

"Something unquenched, unquenchable, is in me, that wants to speak out. A craving for love is in me..." Oh yes, finally, yes yes yes. I let her go, and he walked up and reached out his arms and took her face in his hands, and kissed her.

He kissed her again, on her nose and her cheeks, her chin, on the corners of her smiling lips. He kissed her very nicely, too; sweet lingering little pecks. "Who could behold her smile and not dissolve into tears?" he asked me. Ah, yes. Indeed. Tears.

Then he stepped back to admire her, to take in the full measure of his luck. That this, this, should be waiting here, just for him.

She had those long showgirl legs (well, so do I), enhanced breasts (mine are real) set off by feather boas, and she had big unblinking blue eyes that said you were just exactly what she had all her life been waiting for (mine blink). Furthermore, you knew immediately that she would always look at you that way. In fact I think it's safe to say that most of one's first impressions of her were entirely accurate. Her behavior was in no way misleading. She might disappoint you, but she would never let you down.

She was impeccably calm. Here was a woman who could face life's batterings and joys with the same unwavering expression: *Oh, how big and strong you are! J'adore!* She would never complain. She was patient and always willing to listen, but never intrusive. She was never jealous.

Now, I have softer breasts and better legs, and I am warm, I bend and squish and lick and sing, my mouth opens, my hands grip and stroke, I can dance a tango in spike high heels. But I couldn't compete.

Zarathustra whispered to her, something I couldn't hear, then looked into her eyes as if she had agreed, yes, between them there was perfect understanding. Well, perhaps there was. I am only jealous of perfection when I see it taking a man away from me.

"I will rescue you from all corners," he said, gently pushing down her kicking leg. "I will brush dust, spiders, and twilight away from you."

He hoisted her from the platform, knocked over everything, and dislodged most of her feathers. After some fidgeting they fitted together with her stiff welcoming arms hooked over his left shoulder. He grasped her by her sequined ass and carried her toward the exit sign. She smiled at me over his shoulder, apparently delighted, as she was with all things. She went with him into the night with her eyes wide open.

"Come, cold and stiff companion! Let me show you my nocturnal world and the big, round moon and the silver waterfalls by my cave. The dog howls, the moon is shining. Here are caves and thickets: we shall go astray! Give me your hand! Or just a finger! Where now do you take me, you unruly paragon?"

I never saw him again. Though I still hear his voice from time to time, at night, when the ocean mutters

to itself. Perhaps this was to be expected from a man who spent ten years in a cave, whose best friends were a serpent and an eagle. What other things could reach him, on his mountaintop? Even the serpent had to be flown there in the talons of the eagle.

Perhaps, if I had grown wings and thick talons, Zarathustra would have let me carry him away.

Note: All lines spoken by Zarathustra were assembled from Friedrich Nietzsche's Thus Spoke Zarathustra, in the translation by R. J. Hollingdale.

Zombies on the web

Compiled by [David Chalmers](#)



Zombies are hypothetical creatures of the sort that philosophers have been known to cherish. A zombie is physically identical to a normal human being, but completely lacks conscious experience. Zombies look and behave like the conscious beings that we know and love, but "all is dark inside." There is nothing it is like to be a zombie.

Varieties of zombies

There are actually three different kinds of zombies. All of them are like humans in some ways, and all of them are lacking something crucial (something different in each case).

- [Hollywood zombies](#). These are found in zombie B-movies. Their defining feature is that they are dead, but "reanimated". They are typically rather mean, and fond of human flesh. The zombies pictured on this page are mostly Hollywood zombies (though I'm informed that the one at the bottom is really a ghost demon). An expert tells me that the name should be "Pittsburgh zombies", since the most important zombie movies were made in Pittsburgh, but somehow it doesn't have the same ring.
- [Haitian zombies](#). These are found in the voodoo (or vodou) tradition in Haiti. Their defining feature seems to be that they lack free will, and perhaps lack a soul. Haitian zombies were once normal people, but underwent zombification by a "bokor" through spell or potion, and are afterwards used as slaves.
- [Philosophical zombies](#). These are found in philosophical articles on consciousness. Their defining features is that they lack conscious experience, but are behaviorally (and often physically) identical to normal humans.

(There are also zombie [cocktails](#), [Unix processes](#), and [pop groups](#), but three is enough for now. There are still more categories [here](#).)

These three classes are distinct. Hollywood zombies and Haitian zombies are not philosophical zombies, since they typically have behavioral impairments (see [how to identify a zombie](#)) and may have some sensory experiences (e.g., there maybe something it tastes like when a Hollywood zombie eats flesh). Likewise philosophical and Haitian zombies aren't Hollywood zombies, since they don't eat flesh and are arguably alive (though some hold that Haitian zombies are dead). One might make the case that philosophical and Hollywood zombies lack free will and are thus a sort of Haitian zombie, although both claims would be controversial. In any case, many believe that Hollywood zombies are a sort of corruption of Haitian zombies.



Philosophical zombies

It is philosophical zombies that I'm most interested in here, since I'm a philosopher and they raise very interesting issues. The sort I'm most concerned with are zombies that are physically and behaviorally identical to a conscious human, but lack any conscious experience. As in this [case-study](#) of my own zombie twin, for example.

Most people doubt that zombies could exist in the actual world. (In philosophical terms, they are naturally impossible.) But many people think that they are at least logically possible - i.e. that the idea of zombie is internally consistent, and that there is at least a "possible world" where zombies exist. This logical possibility is sometimes used to draw strong conclusions about consciousness (e.g. in my book [The Conscious Mind](#), and elsewhere).

For example:

- It can be used as a way of illustrating the "hard problem" of consciousness: why do physical processes give rise to conscious experience? This question might equally be phrased as "why aren't we zombies?". If any account of physical processes would apply equally well to a zombie world, it is hard to see how such an account can explain the existence of consciousness in our world.
- It can be used to raise questions about the function of consciousness: why did evolution bother to produce us if zombies would have survived and reproduced just as well? (As e.g. [Flanagan and Polger \[1995\]](#) have argued.)

- And it can even be used to argue against materialism. If there is a possible world which is just like this one except that it contains zombies, then that seems to imply that the existence of consciousness is a *further*, nonphysical fact about our world. To put it metaphorically, even after determining the physical facts about our world, God had to "do more work" to ensure that we weren't zombies.

The general point is that the logical possibility of zombies is one way of illustrating that there is no logical *entailment* from physical facts to facts about consciousness, whereas there is such an entailment in most other domains. Of course even the logical possibility of zombies is controversial to some (e.g. [Dennett \[1995\]](#)), as conceivability intuitions are notoriously elusive; and some scientists have been known to wonder whether anything important really follows from what is merely conceivable. I think that most arguments that use zombies can actually be rephrased in a zombie-free way, so that these arguments can be set aside if one prefers; but zombies at least provide a vivid and provocative illustration.



Related ideas

There are two related ideas that turn up elsewhere in the philosophical and psychological literature.

The first is that of a **functional zombie**, a non-conscious system physically different from but functionally isomorphic to a normal human. For example, [a system with silicon chips instead of neurons](#). (This idea also goes by the more prosaic name of "absent qualia".) Some use the logical possibility of such a functional zombie to argue against reductive functionalist theories of consciousness (which hold that consciousness = functioning). Some go further and argue that functional zombies might even exist in the actual world, suggesting that any form of functionalism or artificial intelligence is doomed. Others (like me) deny that functional zombies could actually exist, so that AI is not threatened.

The other related idea is that of the **zombie within**, which has recently gotten some play in psychology and neuroscience. It turns out that quite a lot of human activity can be accomplished unconsciously -- e.g. unconscious perception, memory, and learning. And some (notably [Milner and Goodale](#)) have argued that there are major neural pathways devoted to unconscious processing of visual inputs that leads directly to motor action. This has led some to suggest that each of us contains a "[zombie within](#)" that unconsciously produces many of our motor responses, without our realizing it.

History

As far as I know, the first paper in the philosophical literature to talk at length about zombies under that name was Robert Kirk's "Zombies vs. Materialists" in *Mind* in 1974, although Keith Campbell's 1970 book *Body and Mind* talks about an "imitation-man" which is much the same thing, and the idea arguably goes back to Leibniz's "mill" argument. After Kirk's paper, there was hardly any explicit discussion of zombies in the philosophical literature for a long time (although there was quite a lot on "absent qualia", i.e. functional zombies). When I wrote my 1993 Ph.D. thesis, in which zombies played a central role, there was hardly anything out there. But for one reason or another, zombies have risen from the grave in the last few years; and they turn out to be unaccountably well-represented on the web, in particular.

So here are a few links for the zombieholic.

Papers on zombies

- Selmer Bringsjord (1995), ["In defense of impenetrable zombies"](#)
 - A commentary on Moody 1995 (one can't really distinguish zombies and humans).
- Selmer Bringsjord (1996), ["The zombie attack on the computational conception of mind"](#)
 - Using the possibility of zombies to draw out inconsistencies in Dennett's argument against Searle.
- David Chalmers (1993), ["Self-ascription without qualia: A case-study"](#)
 - A commentary on a paper by Goldman, with a case-study of "Zombie Dave".
- David Chalmers (1995), ["Absent qualia, fading qualia, dancing qualia"](#)
 - Thought-experiments about functional isomorphs, arguing that in the actual world such isomorphs will be conscious.
- David Chalmers (1996), ["The Conscious Mind: In Search of a Fundamental Theory"](#)
 - My book on consciousness, with all the zombie discussions you could ever ask for. Unfortunately most of the zombie stuff isn't available on the web.
- Allin Cottrell (1996), ["Sniffing the camembert: On the conceivability of zombies"](#)
 - A thoughtful paper arguing that zombies as I describe them in my book may not really be conceivable.
- Daniel Dennett (1995), ["The unimagined preposterousness of zombies"](#)
 - Zombies aren't really imaginable
- Daniel Dennett (1999), ["The zombic hunch: Extinction of an intuition?"](#)
 - On why we shouldn't trust our intuitions about zombies
- Owen Flanagan & Tom Polger (1995), ["Zombies and the function of consciousness"](#)
 - Using zombies to question whether consciousness has a function.
- Stevan Harnad (1995), ["Why and how we are not zombies"](#)
 - The epistemic possibility of zombies illustrates the limits of cognitive science.

- Larry Hauser (1995), ["Revenge of the zombies"](#)
 - Uses "good zombies" to fight the "evil zombies" of dualism & identity theory, yielding behaviorism.
- Jaron Lanier (1995), ["You can't argue with a zombie"](#)
 - Dennett & other consciousness skeptics and AI supporters must be zombies.
- Peter Marton (1998), ["Zombies vs. materialists: The battle for conceivability"](#)
 - Zombies are conceivable if and only if materialism is inconceivable.
- Todd Moody (1995), ["Conversations with zombies"](#)
 - Zombies will be distinguishable as they won't worry about consciousness.
- John Perry, ["The zombie argument"](#)
 - Argues that arguments for zombies presuppose epiphenomenalism, and that any reasonable zombie would behave differently from us.
- Tom Polger, [Zombies](#)
 - An overview of many different philosophical issues about zombies.
- Nigel Shardlow, ["Zombies"](#)
 - Zombie behavior would be coincidental or lying; the idea rests on a Cartesian conception of self-knowledge.
- Nigel Thomas (1996), ["Zombie Killer"](#)
 - Zombies would claim to be conscious, and this leads to an incoherence. Whether the claim is interpreted as true, false, or meaningless, trouble for the zombiephile results.

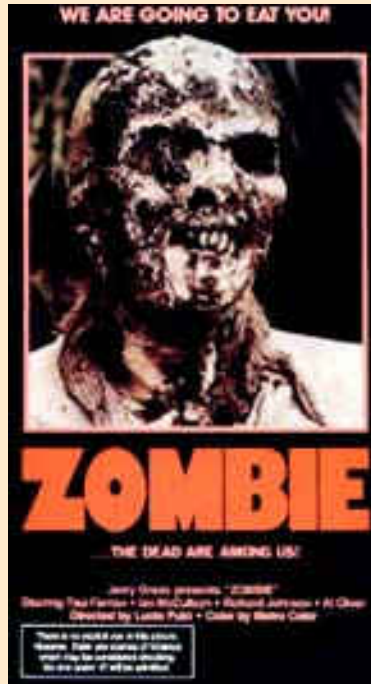
Other philosophical zombie resources

- *Journal of Consciousness Studies* [zombie symposium](#).
- Larry Hauser's [zombie page](#).
- Selmer Bringsjord's [zombie page](#).
- [Zombie discussions](#) on PSYCHE-D (run a find on "zombie")
- [Zombies and modal arguments](#) (from my [philosophy of mind bibliography](#))
- [Experts speak out on the zombie threat](#) ([Zombie Alert](#) commentary on the above papers)
- [The Zombie Within](#) (*New Scientist* special feature)



Zombies of all varieties

- [Zombis](#) (from the [Vodoun/Voodoo information pages](#))
- [First hand accounts of zombification](#) (from [Haiti history archives](#))
- [An overview of Haitian voodoo](#)
- [A dictionary of voodoo terms](#)
- [Zombies and Voodoo Trivia Quiz](#)
- [Zombis may not be what they're reputed to be](#) (and [comment](#))
- [Zombie Activism](#)
- [Zombis and P-Zombis](#)



- [I Walked with a Zombie](#) (and other great [zombie B-movies](#))
- [I Love Zombies!](#)
- [A Guide to Zombie Movies](#) (and [another](#))
- [Zombies-A-Go-Go](#)

- [Survival Guide for a Zombie World](#)
- [Living a Zombie's Dream](#) (confessions of a zombie movie maker)
- [Zombie Biology](#)
- [Field Guide to Zombie Behavior](#)
- [Zombie Gallery](#)



- [Zombies fan page](#) (the British 1960s pop band; and [another page](#))
- [White Zombie](#) (the American 1990s rock group; and [more](#))
- ["Zombie"](#) by the Cranberries
- [Zombie](#) by Joyce Carol Oates
- [Zombification](#) by Andrei Codrescu
- [Zombie children from hell](#) (a short story)
- [Zombie Attack](#) (by a third-grader)
- [Zombie Prom](#) (a musical)
- [Brains4Zombies.com](#) (highly recommended)



- [Unix zombies](#) (undead processes?)
- [Zombie Dinos from Planet Zeltoid](#) (an interactive game)
- [How to make a zombie](#) (the [drink](#); see also this [history with picture](#)).
- [Zombie poker](#) (a card game)
- [Zombie encounter group](#) (the cartoon above, thanks to [Larry Hauser](#))
- [Zombie Alert](#) (zombie warning systems)
- [Experts speak out on the zombie threat!](#) (commentary on this page!)
- And [much, much more](#) (Alta Vista search on "zombi*")



Go to:

- [David Chalmers' home page](#)

Master index to David Chalmers' web site

[Home page](#)

[Papers on consciousness](#)

- [Consciousness and its Place in Nature](#)
- [Facing Up to the Problem of Consciousness](#)
- [Moving Forward on the Problem of Consciousness](#)
- [The Puzzle of Conscious Experience](#)
- [Consciousness and Cognition](#)

- [On the Search for the Neural Correlate of Consciousness](#)
- [What is a Neural Correlate of Consciousness?](#)
- [First-Person Methods in the Science of Consciousness](#)

- [Insentience, Indexicality, and Intensions](#)
- [Self-Ascription Without Qualia: A Case-Study](#)
- [Availability: The Cognitive Basis of Experience?](#)
- [Review of *Journal of Consciousness Studies*](#)
- [Absent Qualia, Fading Qualia, Dancing Qualia](#)
- [What is the Unity of Consciousness?](#)

[Papers on mind, modality, and meaning](#)

- [The Foundations of Two-Dimensional Semantics](#)
- [The Nature of Epistemic Space](#)
- [On Sense and Intension](#)
- [The Components of Content](#)
- [Does Conceivability Entail Possibility?](#)
- [Conceptual Analysis and Reductive Explanation](#)
- [Materialism and the Metaphysics of Modality](#)
- [Mind and Modality](#)
- [The Tyranny of the Subjunctive](#)
- [The Content and Epistemology of Phenomenal Belief](#)

- [The Extended Mind](#)
- [Is There Synonymy in Ockham's Mental Language?](#)

Papers on AI and computation

- [A Computational Foundation for the Study of Cognition](#)
- [Does A Rock Implement Every Finite-State Automaton?](#)
- [Minds, Machines, and Mathematics](#)
- [High-Level Perception, Analogy, and Representation: A Critique of Artificial Intelligence Methodology.](#)
- [Syntactic Transformations on Distributed Representations.](#)
- [Connectionism and Compositionality: Why Fodor and Pylyshyn Were Wrong.](#)
- [The Evolution of Learning: An Experiment in Genetic Connectionism.](#)
- [Subsymbolic Computation and the Chinese Room.](#)

Miscellaneous writings

- [The Two-Envelope Paradox: A Complete Analysis?](#)
- [The St. Petersburg Two-Envelope Paradox](#)
- [The First-Person and Third-Person Views.](#)
- [A Taxonomy of Cognitive Jokes](#)
- [On Spaghetti-Sorters and the Powers of Analog Computation](#)
- [How Cartesian Dualism Might Have Been True](#)
- [Connectionist Representation and Deep Systematicity](#)
- [Determining the Moment of Consciousness?](#)
- [What is it Like to be a Thermostat?](#)
- ["Pick a number between zero and infinity"](#)
- [Is the Continuum Hypothesis true, false, or neither?](#)
- [Thoughts on emergence](#)
- [Does a rock implement every FSA?](#)
- [Realms of cognitive science](#)

The Conscious Mind

- [Table of contents](#)
- [Introduction](#)

- [Jacket copy](#)
- [Precis](#)
- [Reviews](#)
- [Response to Papineau](#)
- [Response to Searle](#)
- [Second response to Searle](#)
- [Response to Mulhauser](#)
- [Response to Hill/McLaughlin, Loar, Shoemaker, and Yablo](#)

Explaining Consciousness: The Hard Problem

- [Table of contents](#)
- [Reviews](#)

Philosophy of Mind: Classical and Contemporary Readings

Contemporary Philosophy of Mind: An Annotated Bibliography

- [Part 1: Consciousness and Qualia](#)
- [Part 2: Mental Content](#)
- [Part 3: Psychophysical Relations and Psychological Explanation](#)
- [Part 4: Philosophy of Artificial Intelligence](#)
- [Part 5: Miscellaneous Topics](#)
- [Part 6: Consciousness in Science](#)

Online Papers on Conscoiusness

- [Part 1: Philosophy of Consciousness](#)
- [Part 2: Other Philosophy of Mind](#)
- [Part 3: Science of Consciousness](#)

Resources

- [People with online papers in philosophy](#)
- [Web resources related to consciousness, philosophy, and such](#)
- [\(Other\) online bibliographies on philosophy, consciousness, and such](#)
- [Philosophical humor](#)

- [Zombies on the web](#)

Classes

- [Philosophy 596V: Foundational Issues in the Science of Consciousness \(fall 1999\)](#)
- [Philosophy 596B: Mind and Modality \(spring 1999\)](#)
- [Philosophy 596B: Meaning, Reason, and Possibility \(spring 2002\)](#)
- [INDV 101: Philosophical Perspectives on the Individual \(fall 1999\)](#)
- [Philosophy 80E: Paradoxes and Dilemmas \(fall 1998\)](#)

Miscellaneous

- [Online discussions of my work](#)
- [Responses to articles on my work](#)
- [Curriculum vitae](#)

[Photo Gallery](#)

- [Tucson hiking](#)
- [Desert life](#)
- [Tucson parties](#)
- [University people](#)

- [Veracruz conference](#)
- [Barcelona conference](#)
- [Tlaxcala conference](#)
- [Syracuse conference](#)
- [Banbury workshop](#)
- [ASSC5 conference](#)
- [Tucson II conference](#)
- [Goldman conference](#)
- [Lehrer celebration](#)

- [Family photos](#)
- [Santa Cruz photos](#)
- [Miscellaneous photos](#)
- [Web photos](#)

Papers on AI and Computation (David Chalmers)

The papers here deal with various aspects of the foundations of artificial intelligence. Some deal with relatively abstract matters about the analysis of computation and its relation to cognition, or about the possibility of artificial intelligence. Others deal with more empirical matters concerning various sorts of artificial intelligence, including symbolic AI, connectionism, and artificial life.

[A Computational Foundation for the Study of Cognition \(PS\)](#)

This paper addresses some key questions about computation and its role in cognitive science. I give an account of what it takes for a physical system to implement a given computation (in terms of abstract patterns of causal organization), and use this account to defend "strong artificial intelligence" and justify the centrality of computational explanation in cognitive science. This paper is so far unpublished, though parts of it have appeared as "On Implementing a Computation" in *Minds and Machines* (1994).

[Does A Rock Implement Every Finite-State Automaton? \(PS\)](#)

In an appendix to his book *Representation and Reality*, Hilary Putnam "proves" that every ordinary open system implements every finite automaton, so that computation cannot provide a nonvacuous foundation for the sciences of the mind. I analyze Putnam's argument and find it wanting. The argument can be patched up to some extent, but this only points the way to a better definition of implementation (of combinatorial-state automata) that is invulnerable to such an objection. A couple of open questions remain, however. This paper appeared in *Synthese* 108: 309-33, 1996.

[Minds, Machines, and Mathematics](#)

This is a commentary on Roger Penrose's book *Shadows of the Mind*, focusing on his attempt to use Gödel's theorem to demonstrate the noncomputability of thought. I argue that the attempt is ultimately unsuccessful, but that there is a novel argument here that many commentators have overlooked, and that it raises many interesting issues. I also comment on his proposals concerning "the missing science of consciousness". This paper appears in [PSYCHE](#), in a [symposium](#) on Penrose's book. Penrose replied in "[Beyond the Doubting of a Shadow](#)".

[High-Level Perception, Analogy, and Representation: A Critique of Artificial Intelligence Methodology](#)

(Co-authored with [Bob French](#) and [Doug Hofstadter](#).) This paper argues that high-level perception is crucially involved in most cognitive processing. We mount a critique of the common approach of using "frozen", hand-coded representations in cognitive modeling (exemplified by Langley and Simon's BACON and Gentner's Structure-Mapping Engine), and argue for a different approach in which

representations are constructed and molded "on the fly". This paper appeared in the [Journal of Experimental and Theoretical Artificial Intelligence](#) in 1992, and also in Hofstadter's book [Fluid Concepts and Creative Analogies](#). You can find a [commentary](#) by Morrison and Dietrich here.

[Syntactic Transformations on Distributed Representations.](#)

In this paper I demonstrate that a connectionist network can be used to perform systematic structure-sensitive transformations on compressed distributed representations of compositional structures. Using representations developed by a [Recursive Auto-Associative Memory](#) (a model of [Jordan Pollack's](#)), a feedforward network learns to move systematically from representations of an active sentence to that of a corresponding passive sentence, and vice versa. This paper appeared in [Connection Science](#) in 1990. This line of research has since been extended by a number of others, e.g. in Lonnie Chrisman's ["Learning Recursive Distributed Representations for Holistic Computation"](#).

[Connectionism and Compositionality: Why Fodor and Pylyshyn Were Wrong.](#)

I point out some structural problems with Fodor and Pylyshyn's arguments against connectionism, and trace these to an underestimation of the role of distributed representation. I discuss some empirical results (from the paper above) that have some bearing on Fodor and Pylyshyn's argument. This paper was published in [Philosophical Psychology](#) in 1993 (an earlier version was in the 1990 *Proceedings of the Cognitive Science Society*). See Murat Aydede's ["Connectionism and the Language of Thought"](#) for some discussion.

[The Evolution of Learning: An Experiment in Genetic Connectionism.](#)

I combine genetic algorithms and neural networks to show how learning mechanisms might evolve in a population of organisms that initially have no capacity to learn. The dynamics of a neural network's cross-time development are specified in a genome, and phenotypes are selected for their ability to learn various tasks across a lifetime. Over many generations, sophisticated learning mechanisms are developed, including on occasion the well-known delta rule. This paper appeared in the *Proceedings of the 1990 Connectionist Summer School Workshop*. See [here](#) for more models of evolution and learning.

[Subsymbolic Computation and the Chinese Room.](#)

In this paper I analyze the distinction between symbolic and subsymbolic computation, and use this to shed some light on Searle's "Chinese Room" argument and the associated argument that "syntax is not sufficient for semantics". I argue that subsymbolic models may be less vulnerable to this argument. I no longer think this paper is very good, but perhaps the analysis of symbolic vs. subsymbolic computation is worthwhile. It appeared in *The Symbolic and Connectionist Paradigms: Closing the Gap*, edited by John Dinsmore, published by Lawrence Erlbaum in 1991.

A Computational Foundation for the Study of Cognition

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

SHORT ABSTRACT

Computation is central to the foundations of modern cognitive science, but its role is controversial. Questions about computation abound: What is it for a physical system to implement a computation? Is computation sufficient for thought? What is the role of computation in a theory of cognition? What is the relation between different sorts of computational theory, such as connectionism and symbolic computation? This article develops a systematic framework that addresses all of these questions. A careful analysis of computation and its relation to cognition suggests that the ambitions of artificial intelligence and the centrality of computation in cognitive science are justified.

LONG ABSTRACT

Computation is central to the foundations of modern cognitive science, but its role is controversial. Questions about computation abound: What is it for a physical system to implement a computation? Is computation sufficient for thought? What is the role of computation in a theory of cognition? What is the relation between different sorts of computational theory, such as connectionism and symbolic computation? In this paper I develop a systematic framework that addresses all of these questions.

Justifying the role of computation requires analysis of *implementation*, the nexus between abstract computations and concrete physical systems. I give such an analysis, based on the idea that a system implements a computation if the causal structure of the system mirrors the formal structure of the computation. This account can be used to justify the central commitments of artificial intelligence and computational cognitive science: the thesis of computational sufficiency, which holds that the right kind of computational structure suffices for the possession of a mind, and the thesis of computational explanation, which holds that computation provides a general framework for the explanation of cognitive processes. The theses are consequences of the facts that (a) computation can specify general patterns of causal organization, and (b) mentality is an *organizational invariant*, rooted in such patterns. Along the way I answer various challenges to the computationalist position, such as those put forward by Searle. I close by advocating a kind of minimal computationalism, compatible with a very wide variety of

empirical approaches to the mind. This allows computation to serve as a true foundation for cognitive science.

Keywords: computation; cognition; implementation; explanation; connectionism; computationalism; representation; artificial intelligence.

1 Introduction

Perhaps no concept is more central to the foundations of modern cognitive science than that of computation. The ambitions of artificial intelligence rest on a computational framework, and in other areas of cognitive science, models of cognitive processes are most frequently cast in computational terms. The foundational role of computation can be expressed in two basic theses. First, underlying the belief in the possibility of artificial intelligence there is a thesis of *computational sufficiency*, stating that the right kind of computational structure suffices for the possession of a mind, and for the possession of a wide variety of mental properties. Second, facilitating the progress of cognitive science more generally there is a thesis of *computational explanation*, stating that computation provides a general framework for the explanation of cognitive processes and of behavior.

These theses are widely held within cognitive science, but they are quite controversial. Some have questioned the thesis of computational sufficiency, arguing that certain human abilities could never be duplicated computationally (Dreyfus 1974; Penrose 1989), or that even if a computation could duplicate human abilities, instantiating the relevant computation would not suffice for the possession of a mind (Searle 1980). Others have questioned the thesis of computational explanation, arguing that computation provides an inappropriate framework for the explanation of cognitive processes (Edelman 1989; Gibson 1979), or even that computational descriptions of a system are vacuous (Searle 1990, 1991).

Advocates of computational cognitive science have done their best to repel these negative critiques, but the positive justification for the foundational theses remains murky at best. Why should *computation*, rather than some other technical notion, play this foundational role? And why should there be the intimate link between computation and cognition that the theses suppose? In this paper, I will develop a framework that can answer these questions and justify the two foundational theses.

In order for the foundation to be stable, the notion of computation itself has to be clarified. The mathematical theory of computation in the abstract is well-understood, but cognitive science and artificial intelligence ultimately deal with physical systems. A bridge between these systems and the abstract theory of computation is required. Specifically, we need a theory of *implementation*: the relation that holds between an abstract computational object (a "computation" for short) and a physical system, such that we can say that in some sense the system "realizes" the computation, and that the computation "describes" the system. We cannot justify the foundational role of computation without first answering the question: *What are the conditions under which a physical system implements a given computation?* Searle (1990) has argued that there is no objective answer to this question, and that any given system can be seen to implement any computation if interpreted appropriately. He argues, for instance, that his wall

can be seen to implement the Wordstar program. I will argue that there is no reason for such pessimism, and that objective conditions can be straightforwardly spelled out.

Once a theory of implementation has been provided, we can use it to answer the second key question: *What is the relationship between computation and cognition?* The answer to this question lies in the fact that the properties of a physical cognitive system that are relevant to its implementing certain computations, as given in the answer to the first question, are precisely those properties in virtue of which (a) the system possesses mental properties and (b) the system's cognitive processes can be explained.

The computational framework developed to answer the first question can therefore be used to justify the theses of computational sufficiency and computational explanation. In addition, I will use this framework to answer various challenges to the centrality of computation, and to clarify some difficult questions about computation and its role in cognitive science. In this way, we can see that the foundations of artificial intelligence and computational cognitive science are solid.

2 A Theory of Implementation

The short answer to question (1) is straightforward. It goes as follows:

A physical system implements a given computation when the causal structure of the physical system mirrors the formal structure of the computation.

In a little more detail, this comes to:

A physical system implements a given computation when there exists a grouping of physical states of the system into state-types and a one-to-one mapping from formal states of the computation to physical state-types, such that formal states related by an abstract state-transition relation are mapped onto physical state-types related by a corresponding causal state-transition relation.

This is still a little vague. To spell it out fully, we must specify the class of computations in question. Computations are generally specified relative to some formalism, and there is a wide variety of formalisms: these include Turing machines, Pascal programs, cellular automata, and neural networks, among others. The story about implementation is similar for each of these; only the details differ. All of these can be subsumed under the class of *combinatorial-state automata* (CSAs), which I will outline shortly, but for the purposes of illustration I will first deal with the special case of *simple finite-state automata* (FSAs).

An FSA is specified by giving a set of input states $I_1, \dots, I_k$, a set of internal states $S_1, \dots, S_m$, and a set of output states $O_1, \dots, O_n$, along with a set of state-transition relations of the form $(S, I) \rightarrow (S', O')$, for each pair (S, I) of internal states and input states, where S' and O' are an internal state and an output state

respectively. S and I can be thought of as the "old" internal state and the input at a given time; S' is the "new" internal state, and O' is the output produced at that time. (There are some variations in the ways this can be spelled out - e.g. one need not include outputs at each time step, and it is common to designate some internal state as a "final" state - but these variations are unimportant for our purposes.) The conditions for the *implementation* of an FSA are the following:

A physical system P implements an FSA M if there is a mapping f that maps internal states of P to internal states of M , inputs to P to input states of M , and outputs of P to output states of M , such that: for every state-transition relation $(S, I) \rightarrow (S', O')$ of M , the following conditional holds: if P is in internal state s and receiving input i where $f(s)=S$ and $f(i)=I$, this reliably causes it to enter internal state s' and produce output o' such that $f(s')=S'$ and $f(o')=O'$. [1]

This definition uses maximally specific physical states s rather than the grouped state-types referred to above. The state-types can be recovered, however: each corresponds to a set $\{s \mid f(s) = S_i\}$, for each $S_i \in M$. From here we can see that the definitions are equivalent. The causal relations between physical state-types will precisely mirror the abstract relations between formal states.

There is a lot of room to play with the details of this definition. For instance, it is generally useful to put restrictions on the way that inputs and outputs to the system map onto inputs and outputs of the FSA. We also need not map *all* possible internal states of P , if some are not reachable from certain initial states. These matters are unimportant here, however. What is important is the overall form of the definition: in particular, the way it ensures that the formal state-transitional structure of the computation mirrors the causal state-transitional structure of the physical system. This is what all definitions of implementation, in any computational formalism, will have in common.

2.1 Combinatorial-state automata

Simple finite-state automata are unsatisfactory for many purposes, due to the monadic nature of their states. The states in most computational formalisms have a combinatorial structure: a cell pattern in a cellular automaton, a combination of tape-state and head-state in a Turing machine, variables and registers in a Pascal program, and so on. All this can be accommodated within the framework of combinatorial-state automata (CSAs), which differ from FSAs only in that an internal state is specified not by a monadic label S , but by a vector $[S^1, S^2, S^3, \dots]$. The elements of this vector can be thought of as the components of the overall state, such as the cells in a cellular automaton or the tape-squares in a Turing machine. There are a finite number of possible values S_j^i for each element S^i , where S_j^i is the j th possible value for the i th element. These values can be thought of as "substates". Inputs and outputs can have a similar sort of complex structure: an input vector is $[I^1, \dots, I^k]$, and so on. State-transition rules are determined by specifying, for each element of the state-vector, a function by which its new state depends on the old overall state-vector and input-vector, and the same for each element of the output-vector.

Input and output vectors are always finite, but the internal state vectors can be either finite or infinite.

The finite case is simpler, and is all that is required for any practical purposes. Even if we are dealing with Turing machines, a Turing machine with a tape limited to 10^{200} squares will certainly be all that is required for simulation or emulation within cognitive science and AI. The infinite case can be spelled out in an analogous fashion, however. The main complication is that restrictions have to be placed on the vectors and dependency rules, so that these do not encode an infinite amount of information. This is not too difficult, but I will not go into details here.

The conditions under which a physical system implements a CSA are analogous to those for an FSA. The main difference is that internal states of the system need to be specified as vectors, where each element of the vector corresponds to an independent element of the physical system. A natural requirement for such a "vectorization" is that each element correspond to a distinct physical region within the system, although there may be other alternatives. The same goes for the complex structure of inputs and outputs. The system implements a given CSA if there exists such a vectorization of states of the system, and a mapping from elements of those vectors onto corresponding elements of the vectors of the CSA, such that the state-transition relations are isomorphic in the obvious way. The details can be filled in straightforwardly, as follows:

A physical system P implements a CSA M if there is a vectorization of internal states of P into components $[s^1, s^2, \dots]$, and a mapping f from the substates s^j into corresponding substates S^j of M , along with similar vectorizations and mappings for inputs and outputs, such that for every state-transition rule $([I^1, \dots, I^k], [S^1, S^2, \dots]) \rightarrow ([S'^1, S'^2, \dots], [O^1, \dots, O^l])$ of M : if P is in internal state $[s^1, s^2, \dots]$ and receiving input $[i^1, \dots, i^n]$ which map to formal state and input $[S^1, S^2, \dots]$ and $[I^1, \dots, I^k]$ respectively, this reliably causes it to enter an internal state and produce an output that map to $[S'^1, S'^2, \dots]$ and $[O^1, \dots, O^l]$ respectively.

Once again, further constraints might be added to this definition for various purposes, and there is much that can be said to flesh out the definition's various parts; a detailed discussion of these technicalities must await another forum (see Chalmers 1996a for a start). This definition is not the last word in a theory of implementation, but it captures the theory's basic form.

One might think that CSAs are not much of an advance on FSAs. Finite CSAs, at least, are no more computationally powerful than FSAs; there is a natural correspondence that associates every finite CSA with an FSA with the same input/output behavior. Of course infinite CSAs (such as Turing machines) are more powerful, but even leaving that reason aside, there are a number of reasons why CSAs are a more suitable formalism for our purposes than FSAs.

First, the *implementation* conditions on a CSA are much more constrained than those of the corresponding FSA. An implementation of a CSA is required to consist in a complex causal interaction among a number of separate parts; a CSA description can therefore capture the causal organization of a system to a much finer grain. Second, the structure in CSA states can be of great *explanatory* utility. A description of a physical system as a CSA will often be much more illuminating than a description as the

corresponding FSA.[2] Third, CSAs reflect in a much more direct way the formal organization of such familiar computational objects as Turing machines, cellular automata, and the like. Finally, the CSA framework allows a unified account of the implementation conditions for both finite and infinite machines.

This definition can straightforwardly be applied to yield implementation conditions for more specific computational formalisms. To develop an account of the implementation-conditions for a Turing machine, say, we need only redescribe the Turing machine as a CSA. The overall state of a Turing machine can be seen as a giant vector, consisting of (a) the internal state of the head, and (b) the state of each square of the tape, where this state in turn is an ordered pair of a symbol and a flag indicating whether the square is occupied by the head (of course only one square can be so occupied; this will be ensured by restrictions on initial state and on state-transition rules). The state-transition rules between vectors can be derived naturally from the quintuples specifying the behavior of the machine-head. As usually understood, Turing machines only take inputs at a single time-step (the start), and do not produce any output separate from the contents of the tape. These restrictions can be overridden in natural ways, for example by adding separate input and output tapes, but even with inputs and outputs limited in this way there is a natural description as a CSA. Given this translation from the Turing machine formalism to the CSA formalism, we can say that a given Turing machine is implemented whenever the corresponding CSA is implemented.

A similar story holds for computations in other formalisms. Some formalisms, such as cellular automata, are even more straightforward. Others, such as Pascal programs, are more complex, but the overall principles are the same. In each case there is some room for maneuver, and perhaps some arbitrary decisions to make (does writing a symbol and moving the head count as two state-transitions or one?) but little rests on the decisions we make. We can also give accounts of implementation for nondeterministic and probabilistic automata, by making simple changes in the definition of a CSA and the corresponding account of implementation. The theory of implementation for combinatorial-state automata provides a basis for the theory of implementation in general.

2.2 Questions answered

The above account may look complex, but the essential idea is very simple: the relation between an implemented computation and an implementing system is one of isomorphism between the formal structure of the former and the causal structure of the latter. In this way, we can see that as far as the theory of implementation is concerned, a computation is simply an *abstract specification of causal organization*. This is important for later purposes. In the meantime, we can now answer various questions and objections.

Does every system implement some computation? Yes. For example, every physical system will implement the simple FSA with a single internal state; most physical systems will implement the 2-state cyclic FSA, and so on. This is no problem, and certainly does not render the account vacuous. That would only be the case if every system implemented *every* computation, and that is not the case.

Does every system implement any given computation? No. The conditions for implementing a given complex computation - say, a CSA whose state-vectors have 1000 elements, with 10 possibilities for each element and complex state-transition relations - will generally be sufficiently rigorous that extremely few physical systems will meet them. What is required is not just a mapping from states of the system onto states of the CSA, as Searle (1990) effectively suggests. The added requirement that the mapped states must satisfy reliable state-transition rules is what does all the work. In this case, there will effectively be at least 10^{1000} constraints on state-transitions (one for each possible state-vector, and more if there are multiple possible inputs). Each constraint will specify one out of at least 10^{1000} possible consequents (one for each possible resultant state-vector, and more if there are outputs). The chance that an arbitrary set of states will satisfy these constraints is something less than one in $(10^{1000})^{10^{1000}}$ (actually significantly less, because of the requirement that transitions be reliable). There is no reason to suppose that the causal structure of an arbitrary system (such as Searle's wall) will satisfy these constraints. It is true that while we lack knowledge of the fundamental constituents of matter, it is impossible to *prove* that arbitrary objects do not implement every computation (perhaps every proton has an infinitely rich internal structure), but anybody who denies this conclusion will need to come up with a remarkably strong argument.

Can a given system implement more than one computation? Yes. Any system implementing some complex computation will simultaneously be implementing many simpler computations - not just 1-state and 2-state FSAs, but computations of some complexity. This is no flaw in the current account; it is precisely what we should expect. The system on my desk is currently implementing all kinds of computations, from EMACS to a clock program, and various sub-computations of these. In general, there is no canonical mapping from a physical object to "the" computation it is performing. We might say that within every physical system, there are numerous computational systems. To this very limited extent, the notion of implementation is "interest-relative". Once again, however, there is no threat of vacuity. The question of whether a given system implements a given computation is still entirely objective. What counts is that a given system does not implement *every* computation, or to put the point differently, that most given computations are only implemented by a very limited class of physical systems. This is what is required for a substantial foundation for AI and cognitive science, and it is what the account I have given provides.

If even digestion is a computation, isn't this vacuous? This objection expresses the feeling that if every process, including such things as digestion and oxidation, implements some computation, then there seems to be nothing special about cognition any more, as computation is so pervasive. This objection rests on a misunderstanding. It is true that any given *instance* of digestion will implement some computation, as any physical system does, but the system's implementing this computation is in general irrelevant to its being an instance of digestion. To see this, we can note that the same computation could have been implemented by various other physical systems (such as my SPARC) without it's being an instance of digestion. Therefore the fact that the system implements the computation is not responsible for the existence of digestion in the system.

With cognition, by contrast, the claim is that it is *in virtue* of implementing some computation that a

system is cognitive. That is, there is a certain class of computations such that *any* system implementing that computation is cognitive. We might go further and argue that every cognitive system implements some computation such that any implementation of the computation would also be cognitive, and would share numerous specific mental properties with the original system. These claims are controversial, of course, and I will be arguing for them in the next section. But note that it is precisely this relation between computation and cognition that gives bite to the computational analysis of cognition. If this relation or something like it did not hold, the computational status of cognition would be analogous to that of digestion.

What about Putnam's argument? Putnam (1988) has suggested that on a definition like this, almost any physical system can be seen to implement every finite-state automaton. He argues for this conclusion by demonstrating that there will almost always be a mapping from physical states of a system to internal states of an FSA, such that over a given time-period (from 12:00 to 12:10 today, say) the transitions between states are just as the machine table say they should be. If the machine table requires that state *A* be followed by state *B*, then every instance of state *A* is followed by state *B* in this time period. Such a mapping will be possible for an inputless FSA under the assumption that physical states do not repeat. We simply map the initial physical state of the system onto an initial formal state of the computation, and map successive states of the system onto successive states of the computation.

However, to suppose that this system implements the FSA in question is to misconstrue the state-transition conditionals in the definition of implementation. What is required is not simply that state *A* be followed by state *B* on all instances in which it happens to come up in a given time-period. There must be a reliable, counterfactual-supporting connection between the states. Given a formal state-transition $A \rightarrow B$, it must be the case that *if* the system were to be in state *A*, it would transit to state *B*. Further, such a conditional must be satisfied for *every* transition in the machine table, not just for those whose antecedent states happen to come up in a given time period. It is easy to see that Putnam's system does not satisfy this much stronger requirement. In effect, Putnam has required only that certain weak material conditionals be satisfied, rather than conditionals with modal force. For this reason, his purported implementations are not implementations at all.

(Two notes. First, Putnam responds briefly to the charge that his system fails to support counterfactuals, but considers a different class of counterfactuals - those of the form "if the system had not been in state *A*, it would not have transited to state *B*". It is not these counterfactuals that are relevant here. Second, it turns out that Putnam's argument for the widespread realization of inputless FSAs can be patched up in a certain way; this just goes to show that inputless FSAs are an inappropriate formalism for cognitive science, due to their complete lack of combinatorial structure. Putnam gives a related argument for the widespread realization of FSAs with input and output, but this argument is strongly vulnerable to an objection like the one above, and cannot be patched up in an analogous way. CSAs are even less vulnerable to this sort of argument. I discuss all this at much greater length in Chalmers 1996a.)

What about semantics? It will be noted that nothing in my account of computation and implementation invokes any semantic considerations, such as the representational content of internal states. This is precisely as it should be: computations are specified syntactically, not semantically. Although it may

very well be the case that any implementations of a given computation share some kind of semantic content, this should be a *consequence* of an account of computation and implementation, rather than built into the definition. If we build semantic considerations into the conditions for implementation, any role that computation can play in providing a foundation for AI and cognitive science will be endangered, as the notion of semantic content is so ill-understood that it desperately needs a foundation itself.

The original account of Turing machines by Turing (1936) certainly had no semantic constraints built in. A Turing machine is defined purely in terms of the mechanisms involved, that is, in terms of syntactic patterns and the way they are transformed. To implement a Turing machine, we need only ensure that this formal structure is reflected in the causal structure of the implementation. Some Turing machines will certainly support a systematic semantic interpretation, in which case their implementations will also, but this plays no part in the definition of what it is to be or to implement a Turing machine. This is made particularly clear if we note that there are some Turing machines, such as machines defined by random sets of state-transition quintuples, that support no non-trivial semantic interpretation. We need an account of what it is to implement these machines, and such an account will then generalize to machines that support a semantic interpretation. Certainly, when computer designers ensure that their machines implement the programs that they are supposed to, they do this by ensuring that the mechanisms have the right causal organization; they are not concerned with semantic content. In the words of Haugeland (1985), if you take care of the syntax, the semantics will take care of itself.

I have said that the notion of computation should not be dependent on that of semantic content; neither do I think that the latter notion should be dependent on the former. Rather, both computation and content should be dependent on the common notion of *causation*. We have seen the first dependence in the account of computation above. The notion of content has also been frequently analyzed in terms of causation (see e.g. Dretske 1981 and Fodor 1987). This common pillar in the analyses of both computation and content allows that the two notions will not sway independently, while at the same time ensuring that neither is dependent on the other for its analysis.

What about computers? Although Searle (1990) talks about what it takes for something to be a "digital computer", I have talked only about computations and eschewed reference to computers. This is deliberate, as it seems to me that computation is the more fundamental notion, and certainly the one that is important for AI and cognitive science. AI and cognitive science certainly do not require that cognitive systems be computers, unless we stipulate that all it takes to be a computer is to implement some computation, in which case the definition is vacuous.

What does it take for something to be a computer? Presumably, a computer cannot merely implement a single computation. It must be capable of implementing many computations - that is, it must be *programmable*. In the extreme case, a computer will be universal, capable of being programmed to compute any recursively enumerable function. Perhaps universality is not required of a computer, but programmability certainly is. To bring computers within the scope of the theory of implementation above, we could require that a computer be a CSA with certain parameters, such that depending on how these parameters are set, a number of different CSAs can be implemented. A universal Turing machine could be seen in this light, for instance, where the parameters correspond to the "program" symbols on

the tape. In any case, such a theory of computers is not required for the study of cognition.

Is the brain a computer in this sense? Arguably. For a start, the brain can be "programmed" to implement various computations by the laborious means of conscious serial rule-following; but this is a fairly incidental ability. On a different level, it might be argued that learning provides a certain kind of programmability and parameter-setting, but this is a sufficiently indirect kind of parameter-setting that it might be argued that it does not qualify. In any case, the question is quite unimportant for our purposes. What counts is that the brain implements various complex computations, not that it is a computer.

3 Computation and cognition

The above is only half the story. We now need to exploit the above account of computation and implementation to outline the relation between computation and cognition, and to justify the foundational role of computation in AI and cognitive science.

Justification of the thesis of computational sufficiency has usually been tenuous. Perhaps the most common move has been an appeal to the Turing test, noting that every implementation of a given computation will have a certain kind of behavior, and claiming that the right kind of behavior is sufficient for mentality. The Turing test is a weak foundation, however, and one to which AI need not appeal. It may be that any behavioral description can be implemented by systems lacking mentality altogether (such as the giant lookup tables of Block 1981). Even if behavior suffices for *mind*, the demise of logical behaviorism has made it very implausible that it suffices for specific mental properties: two mentally distinct systems can have the same behavioral dispositions. A computational basis for cognition will require a tighter link than this, then.

Instead, the central property of computation on which I will focus is one that we have already noted: the fact that a computation provides an abstract specification of the causal organization of a system. Causal organization is the nexus between computation and cognition. If cognitive systems have their mental properties in virtue of their causal organization, and if that causal organization can be specified computationally, then the thesis of computational sufficiency is established. Similarly, if it is the causal organization of a system that is primarily relevant in the explanation of behavior, then the thesis of computational explanation will be established. By the account above, we will always be able to provide a computational specification of the relevant causal organization, and therefore of the properties on which cognition rests.

3.1 Organizational invariance

To spell out this story in more detail, I will introduce the notion of the *causal topology* of a system. The causal topology represents the abstract causal organization of the system: that is, the pattern of interaction among parts of the system, abstracted away from the make-up of individual parts and from the way the causal connections are implemented. Causal topology can be thought of as a dynamic

topology analogous to the static topology of a graph or a network. Any system will have causal topology at a number of different levels. For the cognitive systems with which we will be concerned, the relevant level of causal topology will be a level fine enough to determine the causation of behavior. For the brain, this is probably the neural level or higher, depending on just how the brain's cognitive mechanisms function. (The notion of causal topology is necessarily informal for now; I will discuss its formalization below.)

Call a property P an *organizational invariant* if it is invariant with respect to causal topology: that is, if any change to the system that preserves the causal topology preserves P . The sort of changes in question include: (a) moving the system in space; (b) stretching, distorting, expanding and contracting the system; (c) replacing sufficiently small parts of the system with parts that perform the same local function (e.g. replacing a neuron with a silicon chip with the same I/O properties); (d) replacing the causal links between parts of a system with other links that preserve the same pattern of dependencies (e.g., we might replace a mechanical link in a telephone exchange with an electrical link); and (e) any other changes that do not alter the pattern of causal interaction among parts of the system.

Most properties are not organizational invariants. The property of flying is not, for instance: we can move an airplane to the ground while preserving its causal topology, and it will no longer be flying. Digestion is not: if we gradually replace the parts involved in digestion with pieces of metal, while preserving causal patterns, after a while it will no longer be an instance of digestion: no food groups will be broken down, no energy will be extracted, and so on. The property of being tube of toothpaste is not an organizational invariant: if we deform the tube into a sphere, or replace the toothpaste by peanut butter while preserving causal topology, we no longer have a tube of toothpaste.

In general, most properties depend essentially on certain features that are not features of causal topology. Flying depends on height, digestion depends on a particular physiochemical makeup, tubes of toothpaste depend on shape and physiochemical makeup, and so on. Change the features in question enough and the property in question will change, even though causal topology might be preserved throughout.

3.2 The organizational invariance of mental properties

The central claim of this section is that most mental properties are organizational invariants. It does not matter how we stretch, move about, or replace small parts of a cognitive system: as long as we preserve its causal topology, we will preserve its mental properties.

An exception has to be made for properties that are partly supervenient on states of the environment. Such properties include knowledge (if we move a system that knows that P into an environment where P is not true, then it will no longer know that P), and belief, on some construals where the content of a belief depends on environmental context. However, mental properties that depend only on internal (brain) state will be organizational invariants. This is not to say that causal topology is irrelevant to knowledge and belief. It will still capture the *internal* contribution to those properties - that is, causal topology will contribute as much as the brain contributes. It is just that the environment will also play a

role.

The central claim can be justified by dividing mental properties into two varieties: psychological properties - those that are characterized by their causal role, such as belief, learning, and perception - and phenomenal properties, or those that are characterized by way in which they are consciously experienced. Psychological properties are concerned with the sort of thing the mind *does*, and phenomenal properties are concerned with the way it *feels*. (Some will hold that properties such as belief should be assimilated to the second rather than the first class; I do not think that this is correct, but nothing will depend on that here.)

Psychological properties, as has been argued by Armstrong (1968) and Lewis (1972) among others, are effectively defined by their role within an overall causal system: it is the pattern of interaction between different states that is definitive of a system's psychological properties. Systems with the same causal topology will share these patterns of causal interactions among states, and therefore, by the analysis of Lewis (1972), will share their psychological properties (as long as their relation to the environment is appropriate).

Phenomenal properties are more problematic. It seems unlikely that these can be *defined* by their causal roles (although many, including Lewis and Armstrong, think they might be). To be a conscious experience is not to perform some role, but to have a particular feel. These properties are characterized by *what it is like* to have them, in Nagel's (1974) phrase. Phenomenal properties are still quite mysterious and ill-understood.

Nevertheless, I believe that they can be seen to be organizational invariants, as I have argued elsewhere. The argument for this, very briefly, is a *reductio*. Assume conscious experience is not organizationally invariant. Then there exist systems with the same causal topology but different conscious experiences. Let us say this is because the systems are made of different materials, such as neurons and silicon; a similar argument can be given for other sorts of differences. As the two systems have the same causal topology, we can (in principle) transform the first system into the second by making only gradual changes, such as by replacing neurons one at a time with I/O equivalent silicon chips, where the overall pattern of interaction remains the same throughout. Along the spectrum of intermediate systems, there must be two systems between which we replace less than ten percent of the system, but whose conscious experiences differ. Consider these two systems, *N* and *S*, which are identical except in that some circuit in one is neural and in the other is silicon.

The key step in the thought-experiment is to take the relevant neural circuit in *N*, and to install alongside it a causally isomorphic silicon back-up circuit, with a switch between the two circuits. What happens when we flip the switch? By hypothesis, the system's conscious experiences will change: say, for purposes of illustration, from a bright red experience to a bright blue experience (or to a faded red experience, or whatever). This follows from the fact that the system after the change is a version of *S*, whereas before the change it is just *N*.

But given the assumptions, there is no way for the system to *notice* these changes. Its causal topology stays constant, so that all of its functional states and behavioral dispositions stay fixed. If noticing is defined functionally (as it should be), then there is no room for any noticing to take place, and if it is not, any noticing here would seem to be a thin event indeed. There is certainly no room for a thought "Hmm! Something strange just happened!", unless it is floating free in some Cartesian realm.[3] Even if there were such a thought, it would be utterly impotent; it could lead to no change of processing within the system, which could not even mention it. (If the substitution were to yield some change in processing, then the systems would not have the same causal topology after all. Recall that the argument has the form of a *reductio*.) We might even flip the switch a number of times, so that red and blue experiences "dance" before the system's inner eye; it will never notice. This, I take it, is a *reductio ad absurdum* of the original hypothesis: if one's experiences change, one can potentially notice in a way that makes some causal difference. Therefore the original assumption is false, and phenomenal properties are organizational invariants. This needs to be worked out in more detail, of course. I give the details of this "Dancing Qualia" argument along with a related "Fading Qualia" argument in (Chalmers 1995).

If all this works, it establishes that most mental properties are organizational invariants: any two systems that share their fine-grained causal topology will share their mental properties, modulo the contribution of the environment.

3.3 Justifying the theses

To establish the thesis of computational sufficiency, all we need to do now is establish that organizational invariants are fixed by some computational structure. This is quite straightforward.

An organizationally invariant property depends only on some pattern of causal interaction between parts of the system. Given such a pattern, we can straightforwardly abstract it into a CSA description: the parts of the system will correspond to elements of the CSA state-vector, and the patterns of interaction will be expressed in the state-transition rules. This will work straightforwardly as long as each part has only a finite number of states that are relevant to the causal dependencies between parts, which is likely to be the case in any biological system whose functions cannot realistically depend on infinite precision. (I discuss the issue of analog quantities in more detail below.) Any system that implements this CSA will share the causal topology of the original system. In fact, it turns out that the CSA formalism provides a perfect formalization of the notion of causal topology. A CSA description specifies a division of a system into parts, a space of states for each part, and a pattern of interaction between these states. This is precisely what is constitutive of causal topology.

If what has gone before is correct, this establishes the thesis of computational sufficiency, and therefore the the view that Searle has called "strong artificial intelligence": that there exists some computation such that any implementation of the computation possesses mentality. The fine-grained causal topology of a brain can be specified as a CSA. Any implementation of that CSA will share that causal topology, and therefore will share organizationally invariant mental properties that arise from the brain.

The thesis of computational explanation can be justified in a similar way. As mental properties are organizational invariants, the physical properties on which they depend are properties of causal organization. Insofar as mental properties are to be explained in terms of the physical at all, they can be explained in terms of the causal organization of the system.[4] We can invoke further properties (implementational details) if we like, but there is a clear sense in which they are not vital to the explanation. The neural or electronic composition of an element is irrelevant for many purposes; to be more precise, composition is relevant only insofar as it determines the element's causal role within the system. An element with different physical composition but the same causal role would do just as well. This is not to make the implausible claim that neural properties, say, are entirely irrelevant to explanation. Often the best way to investigate a system's causal organization is to investigate its neural properties. The claim is simply that insofar as neural properties are explanatorily relevant, it is in virtue of the role they play in determining a system's causal organization.

In the explanation of behavior, too, causal organization takes center stage. A system's behavior is determined by its underlying causal organization, and we have seen that the computational framework provides an ideal language in which this organization can be specified. Given a pattern of causal interaction between substates of a system, for instance, there will be a CSA description that captures that pattern. Computational descriptions of this kind provide a general framework for the explanation of behavior.

For some explanatory purposes, we will invoke properties that are not organizational invariants. If we are interested in the biological basis of cognition, we will invoke neural properties. To explain situated cognition, we may invoke properties of the environment. This is fine; the thesis of computational explanation is not an *exclusive* thesis. Still, usually we are interested in neural properties insofar as they determine causal organization, we are interested in properties of the environment insofar as they affect the pattern of processing in a system, and so on. Computation provides a general explanatory framework that these other considerations can supplement.

3.4 Some objections

A computational basis for cognition can be challenged in two ways. The first sort of challenge argues that computation cannot *do* what cognition does: that a computational simulation might not even reproduce human behavioral capacities, for instance, perhaps because the causal structure in human cognition goes beyond what a computational description can provide. The second concedes that computation might capture the capacities, but argues that more is required for true mentality. I will consider four objections of the second variety, and then three of the first. Answers to most of these objections fall directly out of the framework developed above.

But a computational model is just a simulation! According to this objection, due to Searle (1980), Harnad (1989), and many others, we do not expect a computer model of a hurricane to be a real hurricane, so why should a computer model of mind be a real mind? But this is to miss the important point about organizational invariance. A computational simulation is not a mere formal abstraction, but

has rich internal dynamics of its own. If appropriately designed it will share the causal topology of the system that is being modeled, so that the system's organizationally invariant properties will be not merely simulated but *replicated*.

The question about whether a computational model simulates or replicates a given property comes down to the question of whether or not the property is an organizational invariant. The property of being a hurricane is obviously not an organizational invariant, for instance, as it is essential to the very notion of hurricanhood that wind and air be involved. The same goes for properties such as digestion and temperature, for which specific physical elements play a defining role. There is no such obvious objection to the organizational invariance of cognition, so the cases are disanalogous, and indeed, I have argued above that for mental properties, organizational invariance actually holds. It follows that a model that is computationally equivalent to a mind will itself be a mind.

Syntax and semantics. Searle (1984) has argued along the following lines: (1) A computer program is syntactic; (2) Syntax is not sufficient for semantics; (3) Minds have semantics; therefore (4) Implementing a computer program is insufficient for a mind. Leaving aside worries about the second premise, we can note that this argument equivocates between programs and implementations of those programs. While programs themselves are syntactic objects, implementations are not: they are real physical systems with complex causal organization, with real physical causation going on inside. In an electronic computer, for instance, circuits and voltages push each other around in a manner analogous to that in which neurons and activations push each other around. It is precisely in virtue of this causation that implementations may have cognitive and therefore semantic properties.

It is the notion of implementation that does all the work here. A program and its physical implementation should not be regarded as equivalent - they lie on entirely different levels, and have entirely different properties. It is the program that is syntactic; it is the implementation that has semantic content. Of course, there is still a substantial question about how an implementation comes to possess semantic content, just as there is a substantial question about how a *brain* comes to possess semantic content. But once we focus on the implementation, rather than the program, we are at least in the right ball-park. We are talking about a physical system with causal heft, rather than a shadowy syntactic object. If we accept, as is extremely plausible, that brains have semantic properties in virtue of their causal organization and causal relations, then the same will go for implementations. Syntax may not be sufficient for semantics, but the right kind of causation is.

The Chinese room. There is not room here to deal with Searle's famous Chinese room argument in detail. I note, however, that the account I have given supports the "Systems reply", according to which the entire system understands Chinese even if the homunculus doing the simulating does not. Say the overall system is simulating a brain, neuron-by-neuron. Then like any implementation, it will share important causal organization with the brain. In particular, if there is a symbol for every neuron, then the patterns of interaction between slips of paper bearing those symbols will mirror patterns of interaction between neurons in the brain, and so on. This organization is implemented in a baroque way, but we should not let the baroque-ness blind us to the fact that the causal organization - *real*, physical causal organization - is there. (The same goes for a simulation of cognition at level above the neural, in which

the shared causal organization will lie at a coarser level.)

It is precisely in virtue of this causal organization that the system possesses its mental properties. We can rerun a version of the "dancing qualia" argument to see this. In principle, we can move from the brain to the Chinese room simulation in small steps, replacing neurons at each step by little demons doing the same causal work, and then gradually cutting down labor by replacing two neighboring demons by one who does the same work. Eventually we arrive at a system where a single demon is responsible for maintaining the causal organization, without requiring any real neurons at all. This organization might be maintained between marks on paper, or it might even be present inside the demon's own head, if the calculations are memorized. The arguments about organizational invariance all hold here - for the same reasons as before, it is implausible to suppose that the system's experiences will change or disappear.

Performing the thought-experiment this way makes it clear that we should not expect the experiences to be had by the *demon*. The demon is simply a kind of causal facilitator, ensuring that states bear the appropriate causal relations to each other. The conscious experiences will be had by the system as a whole. Even if that system is implemented inside the demon by virtue of the demon's memorization, the system should not be confused with demon itself. We should not suppose that the demon will share the implemented system's experiences, any more than it will share the experiences of an ant that crawls inside its skull: both are cases of two computational systems being implemented within a single physical space. Mental properties arising from distinct computational systems will be quite distinct, and there is no reason to suppose that they overlap.

What about the environment? Some mental properties, such as knowledge and even belief, depend on the environment being a certain way. Computational organization, as I have outlined it, cannot determine the environmental contribution, and therefore cannot fully guarantee this sort of mental property. But this is no problem. All we need computational organization to give us is the *internal* contribution to mental properties: that is, the same contribution that the brain makes (for instance, computational organization will determine the so-called "narrow content" of a belief, if this exists; see Fodor 1987). The full panoply of mental properties might only be determined by computation-plus-environment, just as it is determined by brain-plus-environment. These considerations do not count against the prospects of artificial intelligence, and they affect the aspirations of computational cognitive science no more than they affect the aspirations of neuroscience.

Is cognition computable? In the preceding discussion I have taken for granted that computation can at least *simulate* human cognitive capacity, and have been concerned to argue that this counts as honest-to-goodness mentality. The former point has often been granted by opponents of AI (e.g. Searle 1980) who have directed the fire at the latter, but it is not uncontroversial.

This is to some extent an empirical issue, but the relevant evidence is solidly on the side of computability. We have every reason to believe that the low-level laws of physics are computable. If so, then low-level neurophysiological processes can be computationally simulated; it follows that the function of the whole brain is computable too, as the brain consists in a network of neurophysiological

parts. Some have disputed the premise: for example, Penrose (1989) has speculated that the effects of quantum gravity are noncomputable, and that these effects may play a role in cognitive functioning. He offers no arguments to back up this speculation, however, and there is no evidence of such noncomputability in current physical theory (see Pour-El and Richards (1989) for a discussion). Failing such a radical development as the discovery that the fundamental laws of nature are uncomputable, we have every reason to believe that human cognition can be computationally modeled.

What about Gödel's theorem? Gödel's theorem states that for any consistent formal system, there are statements of arithmetic that are unprovable within the system. This has led some (Lucas 1963; Penrose 1989) to conclude that humans have abilities that cannot be duplicated by any computational system. For example, our ability to "see" the truth of the Gödel sentence of a formal system is argued to be non-algorithmic. I will not deal with this objection in detail here, as the answer to it is not a direct application of the current framework. I will simply note that the assumption that we can see the truth of arbitrary Gödel sentences requires that we have the ability to determine the consistency or inconsistency of any given formal system, and there is no reason to believe that we have this ability in general. (For more on this point, see Putnam 1960, Bowie 1982 and the commentaries on Penrose 1990.)

Discreteness and continuity. An important objection notes that the CSA formalism only captures *discrete* causal organization, and argues that some cognitive properties may depend on continuous aspects of that organization, such as analog values or chaotic dependencies.

A number of responses to this are possible. The first is to note that the current framework can fairly easily be extended to deal with computation over continuous quantities such as real numbers. All that is required is that the various substates of a CSA be represented by a real parameter rather than a discrete parameter, where appropriate restrictions are placed on allowable state-transitions (for instance, we can require that parameters are transformed polynomially, where the requisite transformation can be conditional on sign). See Blum, Shub and Smale (1989) for a careful working-out of some of the relevant theory of computability. A theory of implementation can be given along in a fashion similar to the account I have given above, where continuous quantities in the formalism are required to correspond to continuous physical parameters with an appropriate correspondence in state-transitions.

This formalism is still discrete in time: evolution of the continuous states proceeds in discrete temporal steps. It might be argued that cognitive organization is in fact continuous in time, and that a relevant formalism should capture this. In this case, the specification of discrete state-transitions between states can be replaced by differential equations specifying how continuous quantities change in continuous time, giving a thoroughly continuous computational framework. MacLennan (1990) describes a framework along these lines. Whether such a framework truly qualifies as *computational* is largely a terminological matter, but there it is arguable that the framework is significantly similar in kind to the traditional approach; all that has changed is that discrete states and steps have been "smoothed out".

We need not go this far, however. There are good reasons to suppose that whether or not cognition in the brain is continuous, a discrete framework can capture everything important that is going on. To see this,

we can note that a discrete abstraction can describe and simulate a continuous process to any required degree of accuracy. It might be objected that chaotic processes can amplify microscopic differences to significant levels. Even so, it is implausible that the correct functioning of mental processes *depends* on the precise value of the tenth decimal place of analog quantities. The presence of background noise and randomness in biological systems implies that such precision would inevitably be "washed out" in practice. It follows that although a discrete simulation may not yield precisely the behavior that a given cognitive system produces on a given occasion, it will yield plausible behavior that the system *might* have produced had background noise been a little different. This is all that a proponent of artificial intelligence need claim.

Indeed, the presence of noise in physical systems suggests that any given continuous computation of the above kinds can never be reliably implemented in practice, but only approximately implemented. For the purposes of artificial intelligence we will do just as well with discrete systems, which can also give us approximate implementations of continuous computations.

It follows that these considerations do not count against the theses of computational sufficiency or of computational explanation. To see the first, note that a discrete simulation can replicate everything *essential* to cognitive functioning, for the reasons above, even though it may not duplicate every last detail of a given episode of cognition. To see the second, note that for similar reasons the precise values of analog quantities cannot be relevant to the explanation of our cognitive *capacities*, and that a discrete description can do the job.

This is not to exclude continuous formalisms from cognitive explanation. The thesis of computational explanation is not an exclusive thesis. It may be that continuous formalisms will provide a simpler and more natural framework for the explanation of many dynamic processes, as we find in the theory of neural networks. Perhaps the most reasonable version of the computationalist view accepts the thesis of (discrete) computational sufficiency, but supplements the thesis of computational explanation with the proviso that continuous computation may sometimes provide a more natural explanatory framework (a discrete explanation could do the same job, but more clumsily). In any case, continuous computation does not give us anything fundamentally new.

4 Other kinds of computationalism

Artificial intelligence and computational cognitive science are committed to a kind of computationalism about the mind, a computationalism defined by the theses of computational sufficiency and computational explanation. In this paper I have tried to justify this computationalism, by spelling out the role of computation as a tool for describing and duplicating causal organization. I think that this kind of computationalism is all that artificial intelligence and computational cognitive science are committed to, and indeed is all that they need. This sort of computationalism provides a *general* framework precisely because it makes so few claims about the *kind* of computation that is central to the explanation and replication of cognition. No matter what the causal organization of cognitive processes turns out to be, there is good reason to believe that it can be captured within a computational framework.

The fields have often been taken to be committed to stronger claims, sometimes by proponents and more often by opponents. For example, Edelman (1989) criticizes the computational approach to the study of the mind on the grounds that:

An analysis of the evolution, development, and structure of brains makes it highly unlikely that they could be Turing machines. This is so because of the enormous individual variation in structure that brains possess at a variety of organizational levels. [...] [Also,] an analysis of both ecological and environmental variation, and of the categorization procedures of animals and humans, makes it highly unlikely that the world (physical and social) can function as a tape for a Turing machine. (Edelman 1989, p. 30.)

But artificial intelligence and computational cognitive science are not committed to the claim that the brain is literally a Turing machine with a moving head and a tape, and even less to the claim that that tape is the environment. The claim is simply that some computational framework can *explain* and *replicate* human cognitive processes. It may turn out that the relevant computational description of these processes is very fine-grained, reflecting extremely complex causal dynamics among neurons, and it may well turn out that there is significant variation in causal organization between individuals. There is nothing here that is incompatible with a computational approach to cognitive science.

In a similar way, a computationalist need not claim that the brain is a von Neumann machine, or has some other specific architecture. Like Turing machines, von Neumann machines are just one kind of architecture, particularly well-suited to programmability, but the claim that the brain implements such an architecture is far ahead of any empirical evidence and is most likely false. The commitments of computationalism are more general.

Computationalism is occasionally associated with the view that cognition is rule-following, but again this is a strong empirical hypothesis that is inessential to the foundations of the fields. It is entirely possible that the only "rules" found in a computational description of thought will be at a very low level, specifying the causal dynamics of neurons, for instance, or perhaps the dynamics of some level between the neural and the cognitive. Even if there are no rules to be found at the cognitive level, a computational approach to the mind can still succeed. Another claim to which a computationalist need not be committed are "the brain is a computer"; as we have seen, it is not computers that are central but computations).

The most ubiquitous "strong" form of computationalism has been what we may call *symbolic computationalism*: the view that cognition is computation over representation (Newell and Simon 1976; Fodor and Pylyshyn 1988). To a first approximation, we can cash out this view as the claim that the *computational* primitives in a computational description of cognition are also *representational* primitives. That is to say, the basic syntactic entities between which state-transitions are defined are themselves bearers of semantic content, and are therefore *symbols*.

Symbolic computationalism has been a popular and fruitful approach to the mind, but it does not exhaust the resources of computation. Not all computations are symbolic computations. We have seen that there

are some Turing machines that lack semantic content altogether, for instance. Perhaps systems that carry semantic content are more plausible models of cognition, but even in these systems there is no reason why the content must be carried by the systems' computational primitives. In connectionist systems, for example, the basic bearers of semantic content are *distributed* representations, patterns of activity over many units, whereas the computational primitives are simple units that may themselves lack semantic content. To use Smolensky's term (Smolensky 1988), these systems perform *subsymbolic* computation: the level of computation falls below the level of representation.[5] But the systems are computational nevertheless.

Note that the distinction between symbolic and subsymbolic computation does not coincide with the distinction between different computational formalisms, such as Turing machines and neural networks. Rather, the distinction divides the class of computations within each of these formalisms. Some Turing machines perform symbolic computation, and some perform subsymbolic computation; the same goes for neural networks. (Of course it is sometimes said that all Turing machines perform "symbol manipulation", but this holds only if the ambiguous term "symbol" is used in a purely syntactic sense, rather than in the semantic sense I am using here.)

Both proponents and opponents of a computational approach have often implicitly identified computation with symbolic computation. A critique called *What Computers Can't Do* (Dreyfus 1972), for instance, turns out to be largely directed at systems that perform computation over explicit representation. Other sorts of computation are left untouched, and indeed systems performing subsymbolic computation seem well-suited for some of Dreyfus's problem areas. The broader ambitions of artificial intelligence are therefore left intact.

On the other side of the fence, Fodor (1992) uses the name "Computational Theory of Mind" for a version of symbolic computationalism, and suggests that Turing's main contribution to cognitive science is the idea that syntactic state-transitions between symbols can be made to respect their semantic content. This strikes me as false. Turing was concerned very little with the semantic content of internal states, and the concentration on symbolic computation came later. Rather, Turing's key contribution was the formalization of the notion of *mechanism*, along with the associated *universality* of the formalization. It is this universality that gives us good reason to suppose that computation can do almost anything that any mechanism can do, thus accounting for the centrality of computation in the study of cognition.

Indeed, a focus on symbolic computation sacrifices the universality that is at the heart of Turing's contribution. Universality applies to entire classes of automata, such as Turing machines, where these classes are defined syntactically. The requirement that an automaton performs computation over representation is a strong further constraint, a semantic constraint that plays no part in the basic theory of computation. There is no reason to suppose that the much narrower class of Turing machines that perform symbolic computation is universal. If we wish to appeal to universality in a defense of computationalism, we must cast the net more widely than this.[6]

The various strong forms of computationalism outlined here are bold empirical hypotheses with varying

degrees of plausibility. I suspect that they are all false, but in any case their truth and falsity is not the issue here. Because they are such strong empirical hypotheses, they are in no position to serve as a *foundation* for artificial intelligence and computational cognitive science. If the fields were committed to these hypotheses, their status would be much more questionable than it currently is. Artificial intelligence and computational cognitive science can survive the discovery that the brain is not a von Neumann machine, or that cognition is not rule-following, or that the brain does not engage in computation over representation, precisely because these are not among the fields' foundational commitments. Computation is much more general than this, and consequently much more robust.[7]

5 Conclusion: Toward a minimal computationalism

The view that I have advocated can be called *minimal computationalism*. It is defined by the twin theses of computational sufficiency and computational explanation, where computation is taken in the broad sense that dates back to Turing. I have argued that these theses are compelling precisely because computation provides a general framework for describing and determining patterns of causal organization, and because mentality is rooted in such patterns. The thesis of computational explanation holds because computation provides a perfect language in which to specify the causal organization of cognitive processes; and the thesis of computational sufficiency holds because in all implementations of the appropriate computations, the causal structure of mentality is replicated.

Unlike the stronger forms of computationalism, minimal computationalism is not a bold empirical hypothesis. To be sure, there are some ways that empirical science might prove it to be false: if it turns out that the fundamental laws of physics are noncomputable and if this noncomputability reflects itself in cognitive functioning, for instance, or if it turns out that our cognitive capacities depend essentially on infinite precision in certain analog quantities, or indeed if it turns out that cognition is mediated by some non-physical substance whose workings are not computable. But these developments seem unlikely; and failing developments like these, computation provides a general framework in which we can express the causal organization of cognition, whatever that organization turns out to be.

Minimal computationalism is compatible with such diverse programs as connectionism, logicism, and approaches focusing on dynamic systems, evolution, and artificial life. It is occasionally said that programs such as connectionism are "noncomputational", but it seems more reasonable to say that the success of such programs would vindicate Turing's dream of a computational intelligence, rather than destroying it.

Computation is such a valuable tool precisely because almost any theory of cognitive mechanisms can be expressed in computational terms, even though the relevant computational formalisms may vary. All such theories are theories of causal organization, and computation is sufficiently flexible that it can capture almost any kind of organization, whether the causal relations hold between high-level representations or among low-level neural processes. Even such programs as the Gibsonian theory of perception are ultimately compatible with minimal computationalism. If perception turns out to work as the Gibsonians imagine, it will still be mediated by causal mechanisms, and the mechanisms will be

expressible in an appropriate computational form. That expression may look very unlike a traditional computational theory of perception, but it will be computational nevertheless.

In this light, we see that artificial intelligence and computational cognitive science do not rest on shaky empirical hypotheses. Instead, they are consequences of some very plausible principles about the causal basis of cognition, and they are compatible with an extremely wide range of empirical discoveries about the functioning of the mind. It is precisely because of this flexibility that computation serves as a *foundation* for the fields in question, by providing a common framework within which many different theories can be expressed, and by providing a tool with which the theories' causal mechanisms can be instantiated. No matter how cognitive science progresses in the coming years, there is good reason to believe that computation will be at center stage.

References

Armstrong, D.M. 1968. *A Materialist Theory of the Mind*. Routledge and Kegan Paul.

Block, N. 1981. Psychologism and behaviorism. *Philosophical Review* 90:5-43.

Blum, L., Shub, M., and Smale, S. 1989. On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions, and universal machines. *Bulletin (New Series) of the American Mathematical Society* 21(1):1-46.

Bowie, G. 1982. Lucas' number is finally up. *Journal of Philosophical Logic* 11:279-85.

Chalmers, D.J. (1995). Absent qualia, fading qualia, dancing qualia. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schoningh.

Chalmers, D.J. (1996a). Does a rock implement every finite-state automaton? *Synthese*.

Chalmers, D.J. (1996b). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press. Press.

Dietrich, E.S. 1990. Computationalism. *Social Epistemology*.

Dretske, F. 1981. *Knowledge and the Flow of Information*. MIT Press.

Dreyfus, H. 1972. *What Computers Can't Do*. Harper and Row.

Edelman, G.M. 1989. *The Remembered Present: A Biological Theory of Consciousness*. Basic Books.

Fodor, J.A. 1975. *The Language of Thought*. Harvard University Press.

- Fodor, J.A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.
- Fodor, J.A. and Pylyshyn, Z.W. 1988. Connectionism and cognitive architecture. *Cognition* 28:3-71.
- Fodor, J.A. 1992. The big idea: Can there be a science of mind? *Times Literary Supplement* 4567:5-7 (July 3, 1992).
- Gibson, J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.
- Harnad, S. 1989. Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence* 1:5-25.
- Haugeland, J. 1985. *Artificial intelligence: The Very Idea*. MIT Press.
- Lewis, D. 1972. *Psychophysical and theoretical identifications*. *Australasian Journal of Philosophy* 50:249-58.
- Lucas, J.R. 1963. Minds, machines, and Gödel. *Philosophy* 36:112-127.
- MacLennan, B. 1990. Field computation: A theoretical framework for massively parallel analog computation, Parts I - IV. Technical Report CS-90-100. Computer Science Department, University of Tennessee.
- Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 4:435-50.
- Newell, A. and Simon, H.A. 1981. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery* 19:113-26.
- Penrose, R. 1989. *The Emperor's New Mind: Concerning computers, minds, and the laws of physics*. Oxford University Press.
- Penrose, R. 1990. *Precis of The Emperor's New Mind*. *Behavioral and Brain Sciences* 13:643-655.
- Pour-El, M.B., and Richards, J.I. 1989. *Computability in Analysis and Physics*. Springer-Verlag.
- Putnam, H. 1960. Minds and machines. In (S. Hook, ed.) *Dimensions of Mind*. New York University Press.
- Putnam, H. 1967. The nature of mental states. In (W.H. Capitan and D.D. Merrill, eds.) *Art, Mind, and Religion*. University of Pittsburgh Press.

Putnam, H. 1988. *Representation and Reality*. MIT Press.

Pylyshyn, Z.W. 1984. *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press.

Searle, J.R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-57.

Searle, J.R. 1984. *Minds, brains, and science*. Harvard University Press.

Searle, J.R. 1990. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association* 64:21-37.

Searle, J.R. 1991. *The Rediscovery of the Mind*. MIT Press.

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1-23.

Turing, A.M. 1936. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2* 42: 230-65.

Notes

1. I take it that something like this is the "standard" definition of implementation of a finite-state automaton; see, for example, the definition of the description of a system by a probabilistic automaton in Putnam (1967). It is surprising, however, how little space has been devoted to accounts of implementation in the literature in theoretical computer science, philosophy of psychology, and cognitive science, considering how central the notion of computation is to these fields. It is remarkable that there could be a controversy about what it takes for a physical system to implement a computation (e.g. Searle 1990, 1991) at this late date.

2. See Pylyshyn 1984, p. 71, for a related point.

3. In analyzing a related thought-experiment, Searle (1991) suggests that a subject who has undergone silicon replacement might react as follows: "You want to cry out, 'I can't see anything. I'm going totally blind'. But you hear your voice saying in a way that is completely out of your control, 'I see a red object in front of me'" (pp. 66-67). But given that the system's causal topology remains constant, it is very unclear where there is room for such "wanting" to take place, if it is not in some Cartesian realm. Searle suggests some other things that might happen, such as a reduction to total paralysis, but these suggestions require a change in causal topology and are therefore not relevant to the issue of organizational invariance.

4. I am skeptical about whether phenomenal properties can be explained in wholly physical terms. As I

argue in Chalmers (forthcoming c), given any account of the physical or computational processes underlying mentality, the question of why these processes should give rise to conscious experience does not seem to be explainable within physical or computational theory alone. Nevertheless, it remains the case that phenomenal properties *depend* on physical properties, and if what I have said earlier is correct, the physical properties that they depend on are organizational properties. Further, the explanatory gap with respect to conscious experience is compatible with the computational explanation of cognitive processes and of behavior, which is what the thesis of computational explanation requires.

5. Of course there is a sense in which it can be said that connectionist models perform "computation over representation", in that connectionist processing involves the transformation of representations, but this sense is too weak to cut the distinction between symbolic and subsymbolic computation at its joints. Perhaps the most interesting foundational distinction between symbolic and connectionist systems is that in the former but not in the latter, the computational (syntactic) primitives are also the representational (semantic) primitives.

6. It is common for proponents of symbolic computationalism to hold, usually as an unargued premise, that what makes a computation a computation is the fact that it involves representations with semantic content. The books by Fodor (1975) and Pylyshyn (1984), for instance, are both premised on the assumption that there is no computation without representation. Of course this is to some extent a terminological issue, but as I have stressed in 2.2 and here, this assumption has no basis in computational theory and unduly restricts the role that computation plays in the foundations of cognitive science.

7. Some other claims with which computationalism is sometimes associated include "the brain is a computer", "the mind is to the brain as software is to hardware", and "cognition is computation". The first of these is not required, for the reasons given in 2.2: it is not computers that are central to cognitive theory but computations. The second claim is an imperfect expression of the computationalist position for similar reasons: certainly the mind does not seem to be something separable that the brain can load and run, as a computer's hardware can load and run software. Even the third does not seem to me to be central to computationalism: perhaps there is a sense in which it is true, but what is more important is that computation suffices for and explains cognition. See Dietrich (1990) for some related distinctions between computationalism, "computerism", and "cognitivism".

Does a Rock Implement Every Finite-State Automaton?

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

ABSTRACT

Putnam has argued that computational functionalism cannot serve as a foundation for the study of the mind, as every ordinary open physical system implements every finite-state automaton. I argue that Putnam's argument fails, but that it points out the need for a better understanding of the bridge between the theory of computation and the theory of physical systems: the relation of implementation. It also raises questions about the classes of automata that can serve as a basis for understanding the mind. I develop an account of implementation, linked to an appropriate class of automata, such that the requirement that a system implement a given automaton places a very strong constraint on the system. This clears the way for computation to play a central role in the analysis of mind.

1 Introduction

The theory of computation is often thought to underwrite the theory of mind. In cognitive science, it is widely believed that intelligent behavior is enabled by the fact that the mind or the brain implements some abstract automaton: perhaps a Turing machine, a program, an abstract neural network, or a finite-state automaton. The ambitions of artificial intelligence rest on a related claim of computational sufficiency, holding that there is a class of automata such that any implementation of an automaton in that class will possess a mind. A similar claim is often made about many specific mental properties, including properties characteristic of human mentality: that is, it is claimed that there exists a class of automata such that any implementation of an automaton in that class will have the mental property in question. In this way, it is hoped that computation will provide a powerful formalism for the replication and explanation of mentality.

In an appendix to his book *Representation and Reality* (Putnam 1988, pp. 120-125), Hilary Putnam argues for a conclusion that would destroy these ambitions. Specifically, he claims that *every ordinary open system realizes every abstract finite automaton*. He puts this forward as a theorem, and offers a detailed proof. If this is right, a simple system such as a rock implements any automaton one might imagine. Together with the thesis of computational sufficiency, this would imply that a rock has a mind, and

possesses many properties characteristic of human mentality. If Putnam's result is correct, then, we must either embrace an extreme form of panpsychism or reject the principle on which the hopes of artificial intelligence rest. Putnam himself takes the result to show that computational functionalism cannot provide a foundation for a theory of mind. [[Searle (1990) argues for a similar conclusion, although his argument is much less detailed than Putnam's. I comment on it briefly toward the end of the paper.]]

In what follows I will argue that Putnam's argument does not establish such a conclusion, and that the foundations of artificial intelligence are not in danger. His argument nevertheless points to the need for a better understanding of implementation, the relation that connects the theory of computation with the theory of physical systems. (Putnam calls this relation 'realization'.) An analysis of Putnam's argument leads us to a better understanding of implementation, as well as to insights into the class of automata that are relevant in computational theories of cognition. As objections are raised and modifications are made, the discussion that follows will touch on a number of formalisms in the theory of computation and various constraints on implementation, before settling on a formalism and an account of implementation that avoid the most serious problems.

2 Putnam's Argument

Putnam precedes his proof with discussion of some physical principles. There is a Principle of Physical Continuity, which is not important for our purposes (it plays a minor role in the argument in an application I will accept), and a more central Principle of Noncyclical Behavior. The latter principle holds that every ordinary system is in different maximal states at different times. (A "maximal" state is a total state of the system, specifying the system's physical makeup in perfect detail). Putnam argues for this on the basis that every such system is exposed to electromagnetic and gravitational radiation from a natural clock. I will accept the principle in the discussion that follows. Even if it does not hold across the board (arguably, signals from a number of sources might cancel each other's effects, leading to a cycle in behavior), the more limited result that every noncyclic system implements every finite-state automaton would still be a strong one.

Putnam is centrally concerned with finite-state automata, or FSAs, and in particular FSAs without input or output, or *inputless FSAs*. (He extends his results to FSAs with inputs and outputs later.) Such an FSA is specified by a set of formal states $\{S_1, \dots, S_n\}$, and by a set of state-transition relations which specify for each state the state that must follow. He does not offer an explicit account of the conditions under which a physical system implements an FSA, but from his discussion it is clear that he requires something like the following.

A physical system implements an inputless FSA in a given time-period if there is a mapping f from physical states of the system to formal states of the FSA such that: if the system is in physical state p during the time-period, this causes it to transit into a state q such that formal state $f(p)$ transits to formal state $f(q)$ in the specification of the FSA.

I will later argue that this account is not quite right, but it captures the central facet of the notion of

implementation: the requirement that in order for a physical system to implement a formal automaton, causal relations among states of a physical system must mirror formal relations among states of the automaton.

Putnam considers an arbitrary system S - say, a rock - over an arbitrary time period, such as from 12:00 to 12:07. Using the expression $St(S,t)$ to denote the maximal state of S at t , he defines states of S in the following way.

Let the beginnings of the intervals during which S is to be in one of its stages [...] be $t_1, t_2, \dots, t_n$ (in the example given, $n=7$, and the times in question are $t_1 = 12:00, t_2 = 12:01, t_3 = 12:02, t_4 = 12:03, t_5 = 12:04, t_6 = 12:05, t_7 = 12:06$). The end of the real-time interval during which we wish S to "obey" this table we call $t_{\{n+1\}}$ ($= t_8 = 12:07$, in our example). For each of the states t_i to $t_{\{i+1\}}$, $i = 1, 2, \dots, n$, define a (nonmaximal) *interval state* s_i which is the "region" in phase space consisting of all the maximal states $St(S,t)$ with $t_i \leq t < t_{\{i+1\}}$. (I.e., S is in s_i just in case S is in one of the maximal states in this "region". Note that the system S is in s_1 from t_1 to t_2 , in s_2 from t_2 to t_3 , ..., in s_n from t_n to $t_{\{n+1\}}$. (Left endpoint included in all cases but not the right - this is a convention to ensure the "machine" is in exactly one of the s_i at a given time.) The disjointness of the states s_i is guaranteed by the Principle of Noncyclical Behavior. (pp. 122-123.)

The states are defined, then, so that the system goes through states $s_1, s_2, \dots, s_n$ in succession, in time intervals following $t_1, t_2, \dots, t_n$ respectively.

We can choose an arbitrary (inputless) FSA for the purposes of the argument. Putnam chooses (claiming no loss of generality) an FSA that transits back-and-forth between two states A and B , which is therefore required to go through the sequence $ABABABA$ in the time-period in question.

The argument that the system implements the FSA is straightforward. We simply define physical state a to be the disjunction $s_1 \vee s_3 \vee s_5 \vee s_7$, and state b to be $s_2 \vee s_4 \vee s_6$, and we define the mapping f so that $f(a)=A$ and $f(b)=B$. (Putnam's presentation does not distinguish between physical states a and b and formal states A and B ; I introduce the distinction for clarity.) It is easy to see that during the period in question, the physical system cycles through the states $abababa$, corresponding to formal states $ABABABA$ as required.

Putnam further argues that the system's being in state a *causes* it to transit into state b and vice versa, using the Principle of Continuity to show that given that the system was in state a at a time t with $t_{\{n-1\}} \leq t < t_n$ (and given boundary conditions at t), it was *determined* that the system would be in state b at times t with $t_n \leq t < t_{\{n+1\}}$. I will postpone discussion of this part of the argument.

The argument is general. Given any inputless FSA and any noncyclic physical system, we can map physical states of the system over an arbitrarily long period of time to formal states of the FSA by the

same method. We associate the initial physical state of the system with an initial state of the FSA and associate subsequent physical states of the system with subsequent states of the FSA, where the FSA state-evolution is determined by its state-transition rules. The implementation mapping is determined by taking the disjunction of associated physical states for each state of the FSA, and mapping that disjunctive state to the FSA state in question. Under this mapping, it is easy to see that the evolution of the physical states precisely mirrors evolution of the FSA states. Therefore every noncyclic physical system implements every (inputless) FSA.

3 Analysis

Some object to this argument on the grounds that it requires "unnatural" physical states, involving arbitrary disjunctions. It is difficult to make this objection precise, however, as there does not seem to be an objective distinction between "natural" and "unnatural" states that can do the relevant work; and an objective distinction is needed, if having a mind is to be an objective matter. I will not pursue this line, as I think the problems lie elsewhere. It is also sometimes objected that Putnam employs "time-varying" states, but this is based on a misreading of the argument. The states themselves are quite invariant over time. Time has simply been used to fix reference to these atemporal states.

The problem, I think, is that Putnam's system does not satisfy the right kind of state-transition conditionals. The conditionals involved in the definition of implementation are not ordinary material conditionals, saying that on all those occasions in which the system happens to be in state p in the given time period, state q follows. Rather, these conditionals have modal force, and in particular are required to support counterfactuals: *if* the system were to be in state p , *then* it would transit into state q . This expresses the requirement that the connection between connected states must be *reliable* or *lawful*, and not simply a matter of happenstance. It is required that *however* the system comes to be in state p , it transits into state q (perhaps with some restriction ruling out extraordinary environmental circumstances; perhaps not). We can call this sort of conditional a *strong* conditional.

It is not quite clear whether Putnam intends his conditionals to have this sort of modal force. He requires that the relation be a *causal* relation, and goes to some lengths to argue that his construction indeed satisfies the causal relation by arguing that its being in state a fully determines its transition into state B , so he may have some such requirement in mind. In any case, I will argue that his system does not satisfy the strong conditionals in the way that implementation of an automaton requires. There are two ways in which Putnam's system fails to satisfy the strong conditionals. The first concerns the state-transitions that are actually exhibited. The second concerns unexhibited state-transitions.

To see the first point, consider the transition from state a to state b that the system actually exhibits. For the system to be a true implementation, this transition must be reliable. But Putnam's system is open to all sorts of environmental influences by its very definition as an open system (indeed, Putnam exploits this fact to establish his Principle of Noncyclical Behavior). It follows that if environmental circumstances had been slightly different, the system's behavior would have been quite different. Putnam's construction establishes nothing about the system's behavior under such circumstances. The construction is entirely

specific to the environmental conditions as they were during the time-period in question. It follows that his construction does not satisfy the relevant strong conditionals. Although on this particular run the system happened to transit from state a to state b , if conditions had been slightly different it might have transited from state a to state c , or done something different again.

In particular, it is quite possible for Putnam's system to be in state a at all times between $t_{\{n-1\}}$ and t_n , while failing to be in state b at all times between t_n and $t_{\{n+1\}}$. If environmental conditions had been slightly different at and after time t_n , the system might have been buffeted into an entirely different state well before $t_{\{n+1\}}$. The relevant strong conditionals are therefore not satisfied. It will not do to argue that the system counts as being in state b if it is in that state at the *beginning* of the relevant time-interval. If so, then the same goes for state a , and a similar argument can be given. We only need to note the possibility of environmental buffeting soon after $t_{\{n-1\}}$, so that although the system counts as being in state a in the given interval, it is never in state b in the following interval.

Putnam argues (p. 123) that the system's being in state a throughout the first interval *determines* the transition into state b in the next interval, but his argument falls short of establishing what is required. In particular, he only establishes that:

...given the information that the system was in state A at t , and given that the boundary condition at t was B_t , a mathematically omniscient being can determine from the Principle of Continuity that the system S must have been in $St(S,t)$, and can further determine, given the boundary conditions at subsequent times and the other laws of nature, how S evolves in the whole time interval under consideration. (p. 123)

What is important here is the clause "given boundary conditions at subsequent times". This is illegitimate information for the omniscient being in question to use. What is required is that together with the laws of nature, the state (and perhaps boundary conditions) *during the first interval* determines the state during the next interval. It is illegitimate to appeal to the boundary conditions throughout the next interval. To be sure, if we can independently specify these conditions, then the next state will be entirely fixed, as there will be no room for external buffeting, but this is no more interesting than noting that if we fix the physical state during the next time interval, the physical state will be fixed. This argument therefore fails to establish what is required, and the strong conditional is unsatisfied.

The second failure of the conditionals is perhaps more interesting. To qualify as an implementation, it is not only required that the formal state-transitions that are manifested *on this run* (say, the transitions from A to B and back) are mirrored in the structure of the physical system. It is required that *every* state-transition be so mirrored, including the many that many not be manifested on a particular run. It may be, for instance, that the FSA in question is specified by the state-transitions $A \rightarrow B$, $B \rightarrow A$, $C \rightarrow D$, $D \rightarrow E$, and so on. In this case, we require that the unmanifested features of the machine-table are reflected in the physical structure of the system, so that it will be the case that *if* the system had been in state C , it would have transited to state D , and so on.

Perhaps it might be objected that there is little point having such extra states when the machine eternally

cycles between two states and will never reach the other states. However, an FSA will often specify the machine's behavior for more than one initial state; and more importantly the fact that there is a cycle in the state-space is a contingent feature of Putnam's example, and his argument is supposed to be perfectly general. Generally, a finite-state machine that is of practical interest will not exhaust all its states in a given run, nor will it cycle.

The requirement that unmanifested state-transitions also be reflected in the physical structure of the machine requires a slight change to our definition of implementation. It should read:

A physical system implements an inputless FSA in a given time-period if there is a mapping f from physical states of the system onto formal states of the FSA such that: for every formal state-transition $P \rightarrow Q$ in the specification of the FSA, if the physical system is in a state p such that $f(p)=P$, this causes it to transit into a state q such that $f(q)=Q$.

The changes are (1) the requirement that the mapping map physical states onto *all* states of the FSA, and (2) the requirement that a conditional be satisfied for each state-transition in the FSA. If these conditions are not satisfied, there is a very clear sense in which the FSA is not fully implemented.

It is clear that Putnam's construction does not satisfy this stronger condition. In fact, for any formal state P that does not appear in the sequence manifested in a given run, his construction leaves undefined the physical states that map onto P . Even if we pick arbitrary uninstantiated physical states to map onto P , there is no reason to believe that these will satisfy the requisite state-transition conditionals. (If the conditional in the definition of implementation is read as a *material* conditional, then it turns out that these conditionals are satisfied vacuously - every instance of p transits to q , because there are no instances of p . Again, the conditional needs to be read modally.)

To put the point slightly differently: from the fact that any system that implements a given FSA must go through a sequence *abababa* in real time, the converse does not follow. We cannot infer that any system that goes through the sequence *abababa* implements the FSA. In effect, Putnam has failed to ensure that his system reflects the structure of the FSA. All he has done is ensure that a particular *trace* of the FSA - that is, the states it goes through on a particular run - is reflected by the system's behavior.[*] But much more is required. There are all sorts of inherent possibilities in an FSA that are not reflected in a given trace.

*[[[I owe this observation to Joseph O'Rourke.]]]

In the body of his book (pp. 96-98), Putnam responds briefly to the charge that his construction does not satisfy a certain kind of counterfactual: in particular, the requirement that if the system *had not* been in state a , it would not have transited into state b (this is a requirement on causation according to the analysis in Lewis 1973). His response is that these counterfactuals are dependent on the vague notion of a similarity metric over possible worlds (at least on Lewis' analysis), and that the notions of similarity and possible worlds are even vaguer than what we are trying to explicate.

The counterfactuals that I require do not have this problem. All that my account of implementation requires is that *if* the system is in state *a*, it will transit into state *b*, however it finds itself in the first state. This is in no way a similarity-based counterfactual. It requires that in *all* (or perhaps most) situations in which the system is in state *a*, it will transit into state *b*, no matter how similar or dissimilar that situation is from the actual one. This class of counterfactuals is much less vague than the other kind, and is therefore not open to a similar objection. It simply corresponds to the requirement that the transition be lawful. What makes the law of gravity a law is that under *any* circumstances, if two bodies have an appropriate mass, there will be such-and-such a force exerted. If it only happens between 12:00 and 12:07 today, due to unusual environmental conditions, it does not count.

4 A Possible Reply

It turns out that the above objections can be met fairly easily. To ensure that the physical systems satisfies the relevant strong conditionals as well as the material conditionals, we need only require that the system satisfies a pair of undemanding constraints. It remains the case that problems with strong conditionals undermine Putnam's general thesis, but we will have to wait a while to see why.

To overcome the first objection, it is sufficient to require that the system *reliably* transits through a sequence of states $s_1, s_2, \dots$, irrespective of environmental conditions. (To simplify, I will hereafter assume that time is discrete. Nothing will depend on this.) This is not a difficult requirement: most clocks satisfy it, for instance. Probably most physical systems satisfy such a requirement; perhaps we might find reliable sequences like this in patterns of radiation decay. In any case, let us say a physical system contains a *clock* if it has a subsystem that reliably transits through a sequence like this.

A system containing a clock will circumvent the first objection. If we define the states $s_1, s_2, \dots$ of the system as those states containing the relevant states of the clock, then the transition from s_n to s_{n+1} will be reliable. If disjunctive states *a*, *b*, and so on are defined appropriately, then the transitions between these will satisfy the appropriate strong conditionals.

To reply to the second objection we need to make sure that the system has sufficient extra states to map onto formal states that are not manifested on a given run. We can do this by ensuring that the system contains a *dial*: that is, a subsystem with an arbitrary number of different states, such that when it is put into one of those states it stays in that state come what may. Most physical systems will contain a dial - think of states corresponding to marks on rocks, for instance.

Claim: Every physical system containing a clock and a dial will implement every inputless FSA.

Proof: Label physical states of the system $[i,j]$, where *i* corresponds to a clock state and *j* to a dial state. If the system starts in state $[1,j]$ it will reliably transit through states $[2,j]$, $[3,j]$ and so on.

Assume the physical system on the actual run has dial state 1. The initial state of the system will be $[1,1]$; we associate this state with an initial state of the FSA. We associate states $[2,1]$, $[3,1]$ and so on with

subsequent states of the FSA in the obvious way. If at the end of this process some FSA states have not come up, choose an unmanifested state P , and associate state $[1,2]$ with it. We associate states $[2,2]$, $[3,2]$ and so on with the states that follow P in the evolution of the FSA. If this does not exhaust the states of the FSA, choose an unmanifested state and associate state $[3,1]$ with it, and so on. Eventually all the states of the FSA will be exhausted. For each state of the FSA, there will be a non-empty set of associated physical states $[i,j]$. To obtain the implementation mapping, the disjunction of these states is mapped to the FSA state.

It is easy to see that this system satisfies all the strong conditionals in the strengthened definition of implementation. For every state of the FSA, if the system is (or were to be) in a state that maps onto that formal state, the system will (or would) transit into a state that maps onto the appropriate succeeding formal state. So the result is demonstrated.

There are minor problems with the states $[n,i]$ that come up at the end of the run, where t_n is the final time-step. We need to satisfy conditionals of the form "if it were in state $[n,i]$, it would transit appropriately" - that is, $[n,i]$ would transit to a physical state that maps onto a successor of $f([n,i])$ - but we have established nothing about the physical successors to $[n,i]$. We can get around this problem by assuming the clock has an infinite sequence of states, mapping all the states in each sequence, and mapping an infinite disjunction onto each formal state. This is not an especially unrealistic assumption; there are probably such infinite clocks in most continuous systems.

Thus Putnam's result is preserved in only a slightly weakened form. But there is still no problem for the computationalist about the mind. All this has demonstrated is that inputless FSAs are an inappropriate formalism. The structure of an inputless FSA is quite trivial, and it is entirely the wrong sort of thing to describe or specify a mind. Inputless FSAs are almost never invoked in the theory of computation, precisely because of this triviality.

To see the triviality, note that the state-space of an inputless FSA will consist in a single unbranching sequence of states ending in a cycle, or at best in a finite number of such sequences. The latter possibility arises if there is no state from which every state is reachable. It is possible that the various sequences will join at some point, but this is as far as the "structure" of the state-space goes. This is an completely uninteresting kind of structure, as indeed is witnessed by the fact that it is satisfied by a simple combination of a dial and a clock.

This suggests that we need a more interesting formalism than inputless FSAs. In particular, to put stronger constraints on structure we need to move to FSAs with *input* and perhaps with output. The addition of input changes the formalism from trivial to non-trivial. Where there is input, there can be branching behavior. A formal state can be succeeded by various different formal states, depending on the input. Furthermore, the presence of input gives the formalism a kind of combinatorial structure. Later states depend not just on a single state, but on a combination of state and input.

This formalism is much more appropriate for capturing the dynamics of cognitive systems. Humans do not have a single path of states along which their lives are determined. Even if they do, as some fatalistic

views might suggest, this path does not exhaust their description. For any given sequence of states that a human goes through, it remains the case that *if* things in the world had gone slightly differently, they would have functioned in an interestingly different way. Omitting this potentiality leaves out a vital part of the description of human functioning.[*] A wind-up toy or perhaps a videotape of my life could go through the same sequence of states, but it would not be a cognitive system. Cognition requires at least the possibility of functioning in more than one way.

*[[[Maudlin (1989) raises some very interesting questions about along these lines, questioning the relevance of counterfactual-involving conditionals to conscious experience. For instance, if one blocks the connection that supports some counterfactual conditional in a currently static part of the system, is it plausible that this could change or remove the system's conscious experience? Maudlin suggests that it could not. We have seen, however, that such conditionals are constitutive of computational organization. Using an elaborate construction, Maudlin makes an argument along these lines to the conclusion that cognition is not based in computation. This argument requires an in-depth treatment in its own right.]]]

It might be objected that there are or might be people who have effectively *no* capacity for input and outputs - blind, deaf paralytics for instance -- who nevertheless have minds, and are cognitive systems in a quite reasonable sense. According to this objection, such people cannot be cognitive in virtue of implementing an automaton with input or output, for they have no such input or output themselves.

I think this would be to misdescribe the situation, however. These people can be seen to implement an automaton with input and output. It remains the case when they are in a given state that *if* they were to receive a given input, *then* their state would transit appropriately. It is just that they are currently not receiving any inputs (at best they are receiving a single "null" input) due to the malfunctioning of their sensory apparatus. Of course this means that the inputs in question have to be understood as stimulations of the optic nerve, or of the visual cortex or some other internal system, rather than as patterns of photons (*mutatis mutandis* for other sensory modalities), but this is no problem. It simply moves the boundary inward. The malfunctioning of their sensory system means that their visual cortex does not often get stimulated, but *if* it were to be stimulated, their internal state would change in interesting ways. We can deal with outputs in a similar way.

It is this very sensitivity or potentiality that demonstrates that there is a cognitive system present. If there were only one sequence of states along which a system could "work", it would not be a person but a wind-up toy. The sensitivity to change shows that real cognitive *mechanisms* are present.

An alternative way to handle this sort of case will be outlined later, in terms of automata with internal combinatorial structure. Either way, it turns out that structureless FSAs without input or output are inadequate for the specification of a cognitive system.

5 Finite-State Automata with Input and Output

Putnam extends his result to the case of FSAs with input and output. He notes that there is no hope of

extending the universal realization result to this case. To implement an FSA with input and output, a physical system must at least have the capacity for input and output. Furthermore, it must have something like the right *kind* of input and output, and of input/output dependencies. Putnam notes (p. 124) that inputs and outputs to physical systems cannot map to inputs and outputs of automata arbitrarily, as is the case with internal states. There are generally restrictions on the mapping, depending on our purposes.

Instead, Putnam puts forward a more limited result that is still significant. He argues that a finite-state automaton with input and output is realized by every physical system with the right input/output dependencies. That is, if the physical system gets the input and output right, it gets the FSA right.

To see that this is still significant, note that there are vastly different FSAs with the same input/output pattern. Think of an FSA that outputs zero on every input, for instance, and an FSA that upon receiving an input n goes into an exhaustive calculation to determine whether the number is prime, outputs zero if it is prime, and zero if it is not (it also outputs zero on each step during its calculation). These FSAs have precisely the same input/output dependencies. Nevertheless, one would think that most simple systems that implement the first machine certainly do not implement the second, with its vastly more complex internal structure.

Putnam also notes that if this result is correct, then (computational) *functionalism* about mentality will imply *behaviorism* about mentality. If mentality is dependent on implementing the right automaton, and if an automaton is implemented by any system with the right input and output, then mentality is dependent only on input/output patterns. But functionalism is generally taken to stand in contrast to behaviorism, and to lack many of behaviorism's problems. If Putnam's result is correct, functionalism is no better off. As he summarizes:

Thus we obtain that the assumption that something is a "realization" of a given automaton description (possesses a specified "functional organization") is equivalent to the statement that it behaves as if it had that description. In short, "functionalism", if it were correct, would imply behaviorism! If it is true that to possess mental states is simply to possess a certain "functional organization", then it is also true that to possess mental states is simply to possess certain behavior dispositions! (pp. 124-125)

Putnam's argument for this conclusion is a straightforward variation of his earlier argument. Given an arbitrary FSA, take a system S with the right input/output dependencies. For the usual reasons S will go through a sequence of states s_1, s_2 , and so on. Associate s_1 with the relevant initial state of the FSA, and associate s_2 and subsequent physical states with the relevant subsequent formal states (that is, those FSA states determined by the earlier states and inputs in question). To every FSA state, we map the relevant disjunction of associated physical states. The system therefore goes through relevant states a, b, c that correspond under this mapping to FSA states A, B, C ; these states transit appropriately given the previous state and the input, and appropriate output is produced at each step. So, Putnam argues, the system implements the FSA.

To evaluate this claim, we need to spell out conditions of implementation for an FSA with inputs and

outputs. An FSA with input and output is specified by a set of states, a set of inputs, a set of outputs, and a set of state-transitions $(I,S) \rightarrow (S',O)$ for each input/state pair (I,S) , specifying the next state S' and the output O that will be produced by that state and input. (Perhaps the output should depend only on the previous state and not on the input, but the generalization does not hurt.) We can then straightforwardly extend our previous account of implementation as follows. Of course, we require that the relevant strong conditionals are satisfied.

A physical system implements an FSA (with input and output) in a given time-period if there is a mapping f from physical states of the system onto formal states of the FSA and from inputs and outputs to the system onto inputs and outputs of the FSA such that: for every formal state-transition $(I,S) \rightarrow (S',O)$ in the specification of the FSA, if the physical system is in a state s and receiving input i such that $f(s) = S$ and $f(i) = I$, this causes it to transit into a state s' such that $f(s') = S'$ and to produce output o such that $f(o) = O$.

(There will generally be some further restrictions on the mapping from inputs to inputs and outputs to outputs, but I will not go into this here.)

Does Putnam's construction satisfy this definition? It will have trouble with reliability and with uninstantiated formal states, of course, but as before we can stipulate that it contains a clock and a dial to avoid that problem. There is still a serious problem. We have established that the system transits appropriately in response to the inputs and outputs that it receives. However, we have no guarantee of its transition behavior if the inputs and outputs had been different. Again, in order to satisfy the strong conditionals, it is necessary that *if* it had been in a state/input pair that did not come up in the given run, then it would have transited appropriately.

Nothing in Putnam's construction ensures that this condition is satisfied. In fact there is every reason to believe that it will not be. If the sequence of internal states $s_1, s_2, \dots$ is deterministic or reliable in itself, as it will be in the clock/dial case and as nothing in Putnam's stipulation rules out, then the system will produce precisely the *same* transition behavior for *any* inputs. There is none of the sensitivity that an FSA with input and output is required to have.

Once again, the strong conditionals go unsatisfied. Once again, this is because the structure of the FSA is not fully reflected in the structure of the physical system. All that is represented are a few state-transitions, those that come up on the run in question. If a state-transition in the machine table is not play a role in that run, then it will correspond to nothing in the structure of the physical system. The system therefore fails to implement the FSA.

Because of the combinatorial nature of these state-transitions, it is not as easy to get around this failure as it was before. We cannot simply disjoin states to our heart's content. For every state that takes part in such a disjunction, it is required that for *every* input, it transits appropriately. We cannot simply map the new physical state that follows a given state and input onto the requisite formal state, as it will often be the case that the same physical state can be produced as the result of two distinct state/input pairs. Because of this, the iterative mapping procedure described above will not be well-defined.

Still, this suggests yet another maneuver that might salvage something of Putnam's argument. If we require that for every initial physical state and every sequence of inputs, the system goes into a distinct physical state, the mapping above will be well-defined. Furthermore, this condition is not *too* hard to satisfy. It is satisfied by a system that keeps a record of all its inputs, for instance.

Let us say a physical system contains a *input memory* if it has a subsystem that goes into a distinct state for every possible sequence of inputs (think of it as keeping a list of the inputs so far, perhaps). It then turns out that every physical system with an input memory and a dial implements every FSA with the right input/output dependencies.

To see this, denote states of the physical system by labels $[j, i_1, \dots, i_n]$ (where n can take any value from zero up). The dial state is represented by j , and the sequence of inputs in the recorder is represented by the sequence $i_1, \dots, i_n$. To construct the mapping, assume the dial is in fact set to 1, and associate state $[1]$ with the appropriate state S of the FSA. For every possible input sequence $i_1, \dots, i_n$, we associate physical state $[1, i_1, \dots, i_n]$ with the FSA state that we get by starting at state S and feeding inputs $i_1, \dots, i_n$ (or the formal correlates of these) to the FSA. If there are FSA states that are not reached anywhere in this "tree", we pick such a state, associate it with physical state $[2]$, and start the process again, and repeat until FSA states are exhausted. We map disjunctions of physical states to formal states in the obvious way, and the construction is complete.

It is easy to see that under this mapping, all the relevant strong conditionals are satisfied. If the system is in a physical correlate of a given formal state, and receives some input, it will always transit appropriately. As long as we require that the system has the right input/output *dependencies* (and not just the right input/output patterns in a particular run), then the system will also satisfy the right strong conditionals concerning outputs. It follows that the system implements the automaton.

By requiring the addition of an input memory, we have moved well beyond universal realization, but the result is still troubling. Input memories are not difficult to instantiate. A system that retains a separate mark for every input will achieve it, for instance. It follows for instance that any implementation of a simple zero-outputting FSA, supplemented with a dial and a input memory, will implement the complex primality-testing FSA. This result is quite counterintuitive, but it seems that we are stuck with it.

It also means that if AI is dependent on this FSA formalism, then functionalism still *almost* reduces to behaviorism. We know that everybody with the right behavior will realize a given automaton as long as they carry a videocamera to record inputs and a dial in their pocket. If so, the functionalist premise implies that any two behaviorally equivalent people will be mentally equivalent, as long as they are carrying input memories and dials. The caveat is not much help for non-behavioristic functionalism.

(There is an interesting alternative argument for the reduction of functionalism to behaviorism that Putnam does not mention. There is a plausible general principle about the computational basis of cognition:

(++) If a cognitive system implements two FSAs F and G both of which are sufficient to produce the system's behavioral function, and G implements F , then the cognitive properties of the system arise in virtue of its implementing F (as the added structure of G is just irrelevant implementational detail).

However, by the minimization principle for finite-state automata (Hopcroft and Ullman 1979, p. 67), any two FSAs with the same input/output dependencies implement a common simpler FSA with the same dependencies. Take any two behaviorally equivalent cognitive systems whose cognitive properties arise by virtue of implementing an FSA. By the reduction theorem and (++), these will have the same cognitive properties. It follows that if functionalism is not to reduce to behaviorism, we must either reject (++) or reject FSAs as the basis for cognition.)

It is true that we are no longer in danger of panpsychism, as most systems will not even have the right behavior. However, the result may imply that every FSA will be implemented by a "giant lookup table" (see Block 1981) consisting of a tree with an output listed for each series of inputs. The matter is slightly unclear, as any such tree in our world will be finite; whether a finite tree implements the FSA depends on whether we allow the state-transition conditionals to fail for states at the end of a given time-period (that is, at the outer edge of the tree). But is a human being with a finite lifetime any better off? In any case, even the result that an infinite lookup-table would be as conscious as you and me is troubling.

6 Combinatorial State Automata

It might be tempting to respond to this construction by arguing that somehow these inputs, states, and outputs do not have the right sort of connections between them. After all, there is a reasonable sense in which the state of the recorder is causally irrelevant to the system's function; one might try to introduce some causal relevance condition, or some other requirement on the relation between states. I am not sure that this would work, however; there would be problems with FSAs that always output zero despite complex calculations, for example. I leave the possibility open, because otherwise we are left with the counterintuitive result that the zero-outputter with a dial and a recorder implements the prime-tester, but I will focus here on another way to handle the matter.

The real moral of the above discussion is that even simple FSAs with inputs and outputs are not constrained enough to capture the kind of complex structure that computation and cognition involve. The trouble is that the internal states of these FSAs are *monadic*, lacking any internal structure, whereas the internal states of most computational and cognitive systems have all sorts of complex structure. Generally these states are divisible into components which interact locally and globally according to complex principles. Just as the structure of the system is not captured by a monadic state description, neither are the state-transitions captured by a monadic state-change. There may be all sorts of local dependencies that go into the functioning of such a system.

Turing machines and cellular automata are good examples of computational devices whose states have internal structure. In cognition, it is obvious that the brain has a highly complex structure, with its

functioning consisting in a complex pattern of interaction among billions of neurons and other parts. There is also plausibly complex structure in cognitive processing at coarser levels, structure that is central to the explanation of human thought and behavior.

To capture all this, I will introduce the notion of a *combinatorial state automaton*, or CSA. A CSA is much like an FSA, except that its states have combinatorial structure. An internal state of a CSA is a *vector* $[S^1, S^2, \dots, S^n]$, where the i th component of the vector can take on a finite number of different values, or *substates*. The components of the vectors can be thought of by analogy with squares in a Turing machine or cells in a cellular automaton. The substates correspond to symbols in those squares or particular values for the cells. Inputs and outputs can also have a complex structure if this is desired. State-transition rules are determined by specifying for each component of the state vector a function by which its new substate depends on the old overall state vector and input vector (it will frequently depend on only a few "nearby" components of these vectors), and the same for each element of the output vector. The vectors involved may be finite or infinite, but I will stick to the finite case.

The conditions for implementation of a CSA are analogous to those for an FSA, with appropriate modifications. The main change is that internal states of the physical system must be specified as vectors, where each element of the vector corresponds to an independent element of the physical system. I will stipulate that each element in such a decomposition corresponds to a distinct physical region in the system, although there are other alternatives. The same goes for the complex structure of inputs and outputs, if required. The definition of implementation is as follows:

A physical system implements a given CSA if there is a decomposition of its internal states into substates $[s^1, s^2, \dots, s^n]$, and a mapping f from these substates onto corresponding substates S^j of the CSA, along with similar mappings for inputs and outputs, such that: for every formal state transition $([I^1, \dots, I^k], [S^1, \dots, S^n]) \rightarrow ([S'^1, \dots, S'^n], [O^1, \dots, O^l])$ of the CSA, if the system is in internal state $[s^1, \dots, s^n]$ and receiving input $[i^1, \dots, i^n]$ such that the physical states and inputs map to the formal states and inputs, this causes it to enter an internal state and produce an output that map appropriately to the required formal state and output.

Often the state-transitions of a CSA will be defined in terms of *local* dependencies, as when a substate depends only on a few neighboring substates and perhaps on a few inputs rather than on the entire previous state and input vectors (this will be so for cellular automata and Turing machines, for instance). In this case, we can require that the appropriate restricted conditional holds: that is, if the physical system is in the (few) specified previous substates and receiving the specified inputs, this causes it to transit appropriately.

It will be observed that (finite) CSAs are no more powerful than FSAs. For every CSA, there is an FSA that can simulate it. However, the implementation conditions for CSAs are much more constrained than those for the corresponding FSA. A implementation of a CSA is required to consist in a complex causal interaction among a number of separate parts. Not only do states have to be subdivisible, but the many state-transition conditionals imply that the system must have fine-grained causal structure than an

implementation of the corresponding FSA might well lack.

Because of these constrained implementation conditions, CSAs are much less vulnerable to Putnam-style objections than FSAs. Unlike FSA implementations, CSA implementations are required to have complex internal structure and complex dependencies among their parts. For a complex CSA, the right sort of complex structure will be found in very few physical systems. What does the work is precisely the combinatorial nature of the states, through the requirement that the parts of an internal state are independently variable. This imposes a constrained structure of dependencies on the system that arbitrary systems have no hope of passing.

The CSA formalism also provides another solution to the problem of the blind and deaf paralytic mentioned earlier. A non-trivial computational basis for cognition in this case may be specified by a CSA description, where the CSA need not have inputs and outputs. Even if a CSA lacks inputs and outputs, its combinatorial structure enables it to evade Putnam's objections. We can therefore argue that it is in virtue of implementing this CSA that the person possesses mentality. Certainly such a person will still have complex combinatorial structure in their brain, and there will be all sorts of dependencies such as "*if this neuron were in a different state, then...*". The complexity and sensitivity implied by this formalism can therefore go a long way toward capturing the complex inner life of such a person.

CSAs are a plausible candidate for principles of computational sufficiency, then, due to the fine-grained causal structure they impose on an implementation. They are also much more appropriate for the purposes of cognitive *explanation* than FSAs. Where an FSA description of a process will consist in an unenlightening sequence of monadic states, a CSA description of the same process will provide detailed information about the internal structure and about patterns of interaction between various parts, giving the potential for a much better understanding of the process in question.

CSAs, more than many other computational formalisms, directly reflect certain properties of cognitive dynamics that such a formalism must handle, in order to provide a foundation for a theory of mind. Unlike monadic FSAs but like natural cognitive systems, their internal states have complex structure. Unlike Turing machines and cellular automata (at least as these are most commonly understood), these can take in input and produce output at every time-step, and are therefore quite suitable for the modeling of ongoing situated cognitive function. Nevertheless, the CSA framework is perfectly general. An FSA, a Turing machine, or a cellular automaton can be straightforwardly described as a CSA.

Given an account of implementation for CSAs, it is easy to derive an account of implementation for Turing machines, cellular automata, and so on. We simply have to translate one of these machines into the CSA formalism, which is straightforward. For an implementation of the machine in question, we simply require implementation of the corresponding CSAs. We can derive implementation conditions for programs in languages such as LISP and C by similar methods. Here the translation into the CSA formalism is more complex, but there is no problem in principle.

7 Open questions

The above account of what it is to implement a CSA is useful but not perfect. There are two problems that I can see: the first is that it lets in a few too many systems; the second is that it may let in too few. Neither of these problems are debilitating for the account, but both points suggest the need for a future, more perfect account. Here, I leave both of these as open questions for future work.

1. It turns out that even the CSA framework lets in some Putnam-style false implementations - of a sort. The false implementations here include no systems that we are ever likely to encounter, but they include certain astronomically large systems. Thus we still have counterexamples within the realm of logical possibility, if not within the realm of practical possibility. I give details in what follows.

In attempting to construct Putnam-style counterexamples, one might first try an extension of the "input memory" strategy, by building in an input memory into every component state. For a three-component FSA, for example, one might try an implementation that sends state $[a,b,c]$ with input I into state $[aI, bI, cI]$ (where a, b, c are strings, and aI is a concatenation of a and I); then, we map the strings aI, bI, cI onto the appropriate CSA substates, depending on the substates that a, b , and c mapped to. But it is easy to see that this "implementation" fails to have the requisite combinatorial structure. When we recombine component states here with other component states, we will get the wrong results. For example, *any* state of the form $[a,-,-]$ will transit into one of the form $[aI,-,-]$, but the CSA state-transitions will almost certainly require a different first component in the resulting state depending on the other components in the original state.

For a Putnam-style counterexample to be possible, every component state must be sensitive to every previous component state. The most straightforward way to do this is as follows: build an implementation in which state $[a,b,c]$ with input I transits into state $[abcI, abcI, abcI]$ (where $abcI$ is a concatenation of a, b, c , and I). Now, we are assured that for every resultant component state, there is a unique candidate for the preceding state and input. Then we can construct the natural mapping from strings $abcI$ (in various positions) onto substates of the CSA, without fear of troubles with recombination. A recombined state such as $[a,b',c']$ will transit into a new state with unique component states in every position, each of which can be mapped to the appropriate CSA substate.

But this sensitivity comes at a price. A system like this will suffer from an enormous combinatorial explosion, getting three times bigger at every time-step. If the strings that make up each component have length L at one point, within 100 time-steps they will have length $3^{100}L$, which is about 5.10^{47} times larger. In a very short time, the system will be larger than the known universe! CSAs that are candidates to be bases for cognition will have many more than three components, so the situation there will only be worse. Here, the "implementing" system will reach the boundaries of the universe in number of steps corresponding to a fraction of a second in the life of a brain. So there is no chance that any realistic system could ever qualify as an implementation in this manner.

It is easy to see that any Putnam-style implementation will suffer from such an explosion. To achieve the requisite sensitivity "blindly", every component state must be sensitive to the value of every preceding component state. If a system had step t has n components each with k possible values, then at $t+1$ each

component must be able to carry k^n possible values, at $t+2$ each component must be able to carry k^{n^2} possible values, and so on. Given that encoding k^m values takes a system whose size is proportional to m , we can see that the size of the system must increase by a factor of n at every time step.

No practically possible system has this structure, and it is likely that no physically possible system has it either, as this sort of unbounded explosion is probably incompatible with the physical structure of the universe. It follows that the threat of vacuity and of panpsychism have long since disappeared; the only physically possible systems that implement a non-trivial CSA will do so in virtue of having the right sort of fine-grained causal structure among their components. If the CSA is complex enough to be a plausible basis for mentality, the implementations will themselves be complex enough to be plausible candidates for mentality.

Still, the mere logical possibility of these false implementations is troubling. Intuitively, it seems that the exponentially-increasing system, for all its size, lacks the right sort of specific structure to count as really implementing the corresponding computation. Of course, one might bite the bullet and say that in a system that large, there is enough going on that we can truly find all computations implemented, just as one can find all possible books in Borges' Library of Babel. It might or might not follow that the corresponding minds could be found there too, depending on whether one takes the sufficiency of computational structure for mentality to hold with logical necessity or natural necessity. If it is only a natural necessity, then no conclusions follow from the consideration of a mere logical possibility, but if it is a logical necessity, the stronger conclusion would follow.

On reflection, I think that these systems should be ruled out as implementations, on grounds of lacking the right sort of causal structure. But this means that the account I have given of implementing a CSA is imperfect as it stands. It needs some sort of revision or addition, specifying an additional constraint on causal structure. To find this constraint, we need to reflect upon and analyse the features that the false implementations intuitively lack. One way to cash out the intuitions might be to add a "uniformity" clause, specifying that underlying each formal state-transition of a CSA there needs to be a uniform causal mechanism, as opposed to the highly disjunctive causal mechanism found in the false implementations. Another might be to add a causal relevance clause along the lines suggested before. But it is not obvious just how these suggestions should be spelled out in detail, so I leave the project as a challenge for the future.

In the meantime, we can rest reasonably content with the knowledge that the account as it stands provides satisfactory results within the class of physically possible systems. If a physically possible system implements a CSA on this account, it will do so by virtue of having the right sort of systematic, non-trivial causal dependencies between its component states. But the difficulties in the realm of logical possibility remain something of a challenge.

2. The second worry concerns an area in which the implementation conditions seem not too weak, but too strong. I refer to the constraint that each component in a CSA's state-vector must correspond to a distinct region in the physical system. (This constraint arises from the stipulated definition of a decomposition of a physical system.) One might think that the constraint is somewhat arbitrary. Why cannot components of

a CSA be implemented by overlapping physical states, for example?

Some such independence condition on components is required, however. Otherwise, the conditions for implementing a CSA would collapse into the conditions for implementing the corresponding FSA, and the structure of the CSA would be lost. To see this, consider an arbitrary CSA with an n -element state vector, and the corresponding FSA obtained by collapsing each complex state into a single monadic state, with appropriate state-transitions. Let T be a system that implements the FSA under a mapping f . To see that it implements the CSA under the unrestricted condition: let T^i_j be the substate of the CSA where the i th element takes on its j th value. Let t^i_j be the disjunction of states of T such that $f(t^i_j)$ is an FSA state corresponding to CSA state T^i_j (there will be many such FSA states, of course). Then $t^i_1, t^i_2, \dots$ will be the states of the physical system corresponding to substates $T^i_1, T^i_2, \dots$ of the CSA, and the implementation will hold under the appropriate mapping.

To see what is going on here, note that when the independence condition is left aside, a decomposition of internal states of a physical system into substates is just a set of non-maximal states t^i_j such that the states $t^i_1, t^i_2, \dots$ (representing the various substates of the i th element) are mutually disjoint and together exhaustive of maximal states of the system, for each i . We can think of these as states wherein a parameter has a particular value, with a separate parameter for each element of the vector. If we allow each value of each parameter to be determined by an arbitrary global mapping, we lose the sense that the system is functioning by interaction between independent components, and we therefore lose the extra structure that is vital to the nature of a CSA.

In order to have a true implementation of a CSA, we must require that the various parameters are in some strong sense *independent*, corresponding to separable components of a system. The easiest way to do this is to require that for a system to implement a CSA, the physical differences relevant to the variation in a given parameter - must be restricted to a limited physical region, with different physical regions for each element. That is, the values of a "parameter" supervene on a distinct region for each parameter.

It is possible, however, that a weaker "independence" condition could suffice. After all, not all existing computational systems have disjoint regions for each computational component. Witness systems with virtual memory, where the value of a given CSA "component" can be represented in various parts of the system depending on where there is free space, simply by using a pointer to the correct memory location. Cases such as this one might be handled in terms of more flexible independence conditions on components. An alternative would be to handle such cases by a two-stage definition of implementation, noting that the system implements some more complex CSA under the usual conditions, and the complex CSA bears an appropriate relation to the original CSA. The burden here would rest on the latter relation, but that could be analyzed purely mathematically (in terms of pointer values and the like), avoiding the problems with physical systems and causation.

In any case, the conditions that we have outlined provide at the very least *sufficient* conditions for the implementation of an automaton, admitting a wide class of devices that are appropriate, and excluding counterintuitive "implementations" such as Putnam's. The question of how far these conditions can be loosened without letting in inappropriate systems remains open.

8 Minds, Brains, and Programs

It is worth mentioning an argument by Searle for a thesis similar to Putnam's (Searle 1990). He argues that any physical system can be seen to implement any computation: computational properties are not intrinsic to physics, so that computational descriptions are observer-relative. Under the right interpretation, for example, his wall might be seen to implement the Wordstar program.

From what we have seen here, this argument fails. Whether or not computational properties are intrinsic to physics, the implementation relation between abstract automata and physical systems is perfectly objective. Even if there is a correspondence between states of the wall and states of the Wordstar program, and even if we do not worry about inputs and outputs, it is almost certain that the wall states will not satisfy the relevant strong state-transition conditionals, as the wall will not possess the requisite causal organization. Implementation is a determinate matter, as we have seen, and it puts a very strong constraint on the relevant class of physical systems.

There is a weak sense in which computational descriptions are "observer-relative", however. This reflects the fact that most physical systems will implement more than one abstract automaton. Every system will implement a trivial 1-state CSA, for instance, most systems will implement a 2-state CSA, and so on. So an observer is free to choose between a number of different computational "descriptions" of a system. This poses no problems of vacuity, however. The question of whether or not a given system is accurately "described" by a given computational description remains an objective one. The vacuity problem would arise only if every system implemented *every* CSA, and that is not the case. A given complex CSA will only be implemented by a small fraction of physical systems.

Indeed, it is a commonplace idea that a given physical system can implement more than one abstract computation. A single computer can simultaneously implement a number of programs, for example. One can even have a situation where a program is indirectly implemented, by implementing one program that implements another in turn (where the implementation relation between programs is cashed out in the natural way). It is natural to expect that the brain implements a large number of different automata, depending on what part of the brain we are focusing on and on the level of detail we are concerned with. Descriptions of brain processes in terms of implementations of these automata remain both objective and informative, although which description we appeal to at a given time may depend on our purposes.

There may even be instances in which a single system implements two independent automata, both of which fall into the class sufficient for the embodiment of a mind. A sufficiently powerful future computer running two complex AI programs simultaneously might do this, for example. In this case, we would naturally expect two minds to arise, not one. This suggests that if minds arise in virtue of implementing CSAs, then mental properties should be ascribed to the (system, CSA) pair, rather than to the system alone. Alternatively, it might be more in line with AI terminology to redefine the term "system" to denote the pair in question, in order that a single object can support more than one system. Either way, a given physical hunk of matter can be associated with more than one mind.

This suggests a natural response to Searle's (1980) "Chinese Room" argument. In the Chinese room, where a human is simulating a computer program, there are two distinct CSAs being implemented. The causal organization of the human's brain supports one CSA, and the causal organization between the marks on paper supports another one. We should not let the fact that the human is facilitating the causal dynamics of the second implementation blind ourselves to the fact that the causal structure is there - there are counterfactual-supporting state-transition relations between the marks on paper, just as there are between silicon chips or neurons - and that this structure is quite distinct from the structure in the first CSA implementation. We therefore have a situation analogous to that in which a computer implements two independent programs. There are two distinct causal and computational systems here, and we should expect two distinct minds to be associated with them.

9 Conclusion

Computational descriptions of physical systems need not be vacuous. We have seen that there is a well-motivated formalism, that of combinatorial state automata, and an associated account of implementation, such that the automata in question are implemented approximately when we would expect them to be: when the causal organization of a physical system mirrors the formal organization of an automaton. In this way, we establish a bridge between the formal automata of computation theory and the physical systems of everyday life. We also open the way to a computational foundation for the theory of mind.

This is only half the story, of course, and the easy half. The harder part is to take advantage of this bridge, showing that the physical properties that a computational description fixes are the properties in virtue of which minds arise. It is not implausible that minds arise in virtue of causal organization, but neither is it obvious. It is also plausible but not obvious that the discrete CSA framework can capture the precise causal organization (perhaps continuous, perhaps even non-computable) on which mentality depends. This second half is likely to be a long story, although see Chalmers (1994b) for my own attempt at putting things together. In any case, an account of computation and implementation clears the way.[*]

*[[[The ideas in this paper arose out of discussion and argument with a number of others, largely over the Internet. Thanks are due to David Joslin, Daryl McCullough, Joseph O'Rourke, Calvin Ostrum, and Jerry Seligman.]]]

References

Block, N. 1981. Psychologism and behaviorism. *Philosophical Review* 90:5-43.

Chalmers, D.J. 1994a. On implementing a computation. *Minds and Machines* 4.

Chalmers, D.J. 1994b. A computational foundation for the study of cognition. Philosophy-Neuroscience-Psychology Technical Report 94-03, Washington University.

Hopcroft, J.E., and Ullman, J.D. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Lewis, D. 1973. Causation. *Journal of Philosophy* 70:556-67.

Maudlin, T. 1989. Computation and consciousness. *Journal of Philosophy* 86:407-32.

Putnam, H. 1988. *Representation and Reality*. MIT Press.

Searle, J.R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-57.

Searle, J.R. 1990. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association* 64:21-37.

Minds, Machines, and Mathematics

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

[[Written for a [PSYCHE symposium](#) on Roger Penrose's book *Shadows of the Mind*. Penrose's reply is [here](#).]]

Introduction

In his stimulating book SHADOWS OF THE MIND, Roger Penrose presents arguments, based on Gödel's theorem, for the conclusion that human thought is uncomputable. There are actually two separate arguments in Penrose's book. The second has been widely ignored, but seems to me to be much more interesting and novel than the first. I will address both forms of the argument in some detail. Toward the end, I will also comment on Penrose's proposals for a "new science of consciousness".

1. The First Argument

The best way to address Gödelian arguments against artificial intelligence is to ask: what would we *expect*, given the truth of Gödel's theorem, if our reasoning powers could be captured by some formal system F . One possibility is that F is essentially unsound, so that Gödel's theorem does not apply. But what if F is sound? Then we would expect that:

- (a) F could not prove its Gödel sentence $G(F)$;
- (b) F could prove the conditional "If F is consistent, then $G(F)$ is true";
- (c) F could not prove that F is consistent.

If our reasoning powers are capturable by some sound formal system F , then, we should expect that we will be unable to see that F is consistent. This does not seem too surprising, on the face of it. After all, F is likely to be some extremely complex system, perhaps as complex as the human brain itself, and there is no reason to believe that we can determine the consistency of arbitrary formal systems when those systems are presented to us.

There does not seem to be anything especially paradoxical about this situation. Many arguments from

Gödel's theorem, such as that given by Lucas, founder at just this point: they offer us no reason to believe that we can see the truth of our own Gödel sentence, as we may be unable to see the consistency of the associated formal system. How does Penrose's argument fare?

Penrose is much more cautious in his phrasing. In Chapter 2, he argues carefully for the conclusion that our reasoning powers cannot be captured by a "knowably sound" formal system. This seems to be correct, and indeed mirrors the analysis above. If we are a sound formal system F , we will not be able to determine that F is sound. So far, this offers no threat to the prospects of artificial intelligence. The real burden of Penrose's argument is carried by Chapter 3, then, where he argues that the position that we are a formal system that is not "knowably sound" is untenable.

One position that an advocate of AI might take is to argue that our reasoning is fundamentally unsound, even in an idealization. I will not take this path, however. For a start, I have some sympathy with Penrose's idea that we have an underlying sound competence, even our performance sometimes goes astray. But further, it seems to me that to hold that this is the only problem in Penrose's argument would be to concede too much power to the argument. It would follow, for example, that there are parts of our arithmetical competence that no *sound* formal system could ever duplicate; it would seem that our unsoundness would be essential to our capacity to see the truth of Gödel sentences, for example. This would be a remarkably strong conclusion, and does not seem at all plausible to me. So I think that the deepest problems with Penrose's argument must lie elsewhere.

I will concede to Penrose that we are fundamentally sound, then. As before, the natural position for an advocate of AI is that our powers are captured by some sound formal system F that cannot demonstrate that F is sound. What is Penrose's argument against this position? He has two sub-arguments here, depending on whether we can know that F is the formal system that captures our reasoning.

If we *could* know that F captures our reasoning, Penrose's argument would be very straightforward:

- (1) We know that we are sound;
 - (2) We know that F captures our reasoning;
- so (3) We know that F is sound.

One might question premise (1) - I will raise some problems with it later - but it does have a certain plausible quality. Certainly, it seems antecedently more plausible than the much stronger position that we know that F is sound. But all this is irrelevant, as premise (2) is so implausible. There is very little reason to believe that if our reasoning is captured by F , then we could know that fact.

It might *seem* plausible that we could know that F underlies our processing - why couldn't we just investigate our underlying brain processes? But to do this would be to change the game. It is of no help to Penrose if we can know *using external resources* (such as perception inputs) that F captures our

reasoning. For to use external resources would be to go beyond the resources provided by F itself. And there would be no contradiction in the supposition that F could know, using external resources, that F is consistent, and therefore that $G(F)$ is true. A contradiction would only arise if F could know this wholly under its own steam.

For this argument to be at all relevant, then, we would need to know that F captures our reasoning powers wholly using our internal resources - that is, the resources that F itself provides. But there is not the slightest reason to believe that we could do this. If we are a formal system, we certainly cannot determine which formal system we are on the basis of introspection! So again, the advocate of artificial intelligence is in no danger. She need simply hold the unsurprising position that we are a formal system F , but that we can't tell through introspection that we are F .

To make his case, Penrose needs to argue that if we are a sound formal system F , then we could determine that F is sound, *independently* of any knowledge that we are F . That is, he needs to make the case that if F is presented to us, we could determine that it is sound through an analysis of F alone. This is the burden that Penrose tries to meet in section 3.3. It is this section that effectively carries all the crucial weight; if it does not succeed, then this line of Penrose's argument simply fails.

How does Penrose argue that we could see that F is sound? He argues in 3.3 that we can see F as a system of axioms and inference rules. Clearly, we can see that each of the axioms is true: if F can see their truth, so can we. Further, Penrose argues, we must be able to see that each of the basic inference rules is valid, as it is extremely implausible that our reasoning could rely on inference rules that we regard as "fundamentally dubious". And if we know that the axioms are true and that the inference rules are valid, then we know that F is sound.

But why should we accept that F consists of a set of axioms and inference rules? F , after all, is supposed to potentially correspond to *any* sort of computational system - it might be a simulation of the whole of the human brain, for example. This will not look anything like a neat logical system: we will not be able to decompose it into an underlying set of "axioms" and "rules of procedure". Rather, it will be a big computational system that churns away on a given statement as input, and eventually outputs "yes" or "no".

It is true that for any Turing machine that accepts a certain class of statements, we can find a corresponding axiom-plus-rules system that accepts the same class (or at least the closure of that class under logical consequence). There is a lemma by Craig to this effect; without its applications of Gödel's theorem to draw conclusions about Turing machines would not even get off the ground. But the "axiom-plus-rules" system that we end up with may be extraordinarily complex. In particular, the "inference rules" may be just about as complex as the original system - perhaps equivalent to a complex connectionist procedure for generating further theorems. And as before, there is no reason why we should be able to see that this sort of "rule" should be valid, any more than we could see from an analysis that an overall computational brain process is sound. This is not to say that we think we are relying on "fundamentally dubious" procedures - it is just that the procedures that govern the dynamics of our brain are too complex for us to analyse them as sound or otherwise.

In this section, Penrose seems to assume that the relevant class of computational systems are all something akin to theorem-provers in first-order logic, but of course there is no reason to make such an assumption. For his argument to have its full generality, proving that our physical processes could not even be simulated computationally, it must apply to any sort of computational process. Even within the realm of existing AI research, there are many computational procedures, such as connectionist networks, which are not decomposable into axioms and rules of inference.

(I suspect that even an advocate of logic-based AI might have a response to make here. It might be held, for example, that we may occasionally use certain complex inference rules (when we generate Gödel sentences by transfinite counting, for example), whose validity is not obvious to us on analysis, without this in any way impugning the reliability of our reasoning. We might soundly "use" a procedure despite its resistance to our analysis. This indeed is just what we might expect around the "outer limits" of Gödelization, which after all is really where Penrose's argument gains its force. There is no difficulty in the idea that the reasoning methods we use in *everyday* mathematics can be seen to be sound - Penrose's arguments really apply at the level of our unusual "Gödelizing" procedures, which rely on our ability to count transfinite ordinals. But to be able to see that some Gödelizing rule is valid would be akin to making that last step in a Gödelization procedure, the one that is just complex enough to be beyond us. But I leave these difficult issues aside for now.)

It is section 3.3 that carries the burden of this strand of Penrose's argument, but unfortunately it seems to be one of the least convincing sections in the book. By his assumption that the relevant class of computational systems are all straightforward axiom-and-rules system, Penrose is not taking AI seriously, and certainly is not doing enough to establish his conclusion that physics is uncomputable. I conclude that none of Penrose's argument up to this point put a dent in the natural AI position: that our reasoning powers may be captured by a sound formal system F , where we cannot determine that F is sound.

2. Penrose's Second Argument

Hiding at the back of Chapter 3, however, Penrose has a new argument that escapes many of these problems. It is unfortunate that this argument was so deeply buried; most commentators seem to have missed it. Unlike the previous argument, this argument does *not* depend on the claim that we if we are a sound formal system F , we would be able to see that F is sound. Because of this, it is a more novel and interesting argument, and more worthy of attention.

The argument is developed in a roundabout way (which may have led some readers astray), but is summarized in the fantasy dialogue with a robot mathematician in 3.23. The argument is given in a somewhat indirect form, involving complex procedures by which a given formal system might have evolved, but its basic structure is very simple. In a simplified and somewhat loose form, the argument goes as follows:

(1) Assume my reasoning powers are captured by some formal system F (to put this more briefly, "I am

F)). Consider the class of statements I can know to be true, *given* this assumption.

(2) Given that I know that I am F , I know that F is sound (as I know that I am sound). Indeed, I know that the larger system F' is sound, where F' is F supplemented by the further assumption "I am F ". (Supplementing a sound system with a true statement yields a sound system.)

(3) So I know that $G(F')$ is true, where this is the Gödel sentence of the system F' .

(4) But F' could not see that $G(F')$ is true (by Gödel's theorem).

(5) By assumption, however, I am now effectively equivalent to F' . After all, I am F supplemented by the knowledge that I am F .

(6) This is a contradiction, so the initial assumption must be false, and F must not have captured my powers of reasoning after all.

(7) The conclusion generalizes: my reasoning powers cannot be captured by any formal system.

Strictly speaking, the conclusion that must be drawn is that I cannot *know* that I am identical to a formal system F ; in showing that I can see the truth of $G(F')$, we assumed not just that I am F but that I know I am F . But this is still a strong conclusion. For example, it would rule out even the possibility that we could empirically discover that we were identical to some system F - if we were to "discover" this, the reasoning would lead us to a contradiction. So even this would be threatening to the prospects of AI.

The power of this argument stems from the fact that it does not depend on one's ability to determine that a system F is sound, or to determine that we are F . Rather, it relies on the *assumption* that one is F to reach the relevant conclusions, thus contradicting the assumption. On the face of it one might have thought that making such an assumption would show only that the larger system F' could prove the Gödel sentence of the smaller system F , but the insight of the argument is that things can be bootstrapped into a situation where F' sees its own Gödel sentence, leading to trouble.

As far as I can determine, this argument is free of the obvious flaws that plague other Gödelian arguments, such as Lucas's argument and Penrose's earlier arguments. If it is flawed, the flaws lie deeper. It is true that the argument has a feeling of achieving its conclusion as if by magic. One is tempted to say: "why couldn't F itself engage in just the same reasoning?". But although there are various directions in which one might try to attack the argument, no knockdown refutation immediately presents itself. For this reason, the argument is quite challenging. Compared to previous versions, this argument is much more worthy of attention from supporters of AI.

On reflection, I have come to believe that the greatest vulnerability in this argument lies in the assumption that we know (unassailably) that we are consistent. This assumption seems relatively innocuous, compared to the previous strong claim that we could determine that F is consistent; on the

face of it, it does not seem vastly stronger than the assumption that we are consistent. But I think that in fact, it is this assumption, and not the assumption that we know we are F , that carries the central responsibility for generating the contradiction. I have largely become convinced of this through discussions with Daryl McCullough, and the central argument below (an adaptation of a result of Lob's) was suggested by him.

The best way to see this is to show that the assumption that we know we are consistent *already* leads to a contradiction in its own right, even without the further assumption that we know we are F . Specifically, we can argue that any system that "unassailably" believes in its own consistency will in fact be led to a contradiction (under certain plausible further assumptions). This can be done as follows.

In these matters, we are concerned with a system's reasoning about its own beliefs, as well as about mathematics. So we can assume it has a symbol B , representing belief, where $B(n)$ corresponds to the statement that it believes the statement with Gödel number n . (Below, I abbreviate by writing " $B(A)$ " instead of " $B(\ulcorner A \urcorner)$ ", where $\ulcorner A \urcorner$ is the Gödel number of A .) And let us write $\vdash A$ if the system has the power to "unassailably" assert A . (By using this notation I do not intend to beg the question about whether the system is computational!) Then the following assumptions are reasonable (suppressing universal qualifiers):

(1) If $\vdash A$, then $\vdash B(A)$.

(2) $\vdash B(A_1) \ \& \ B(A_1 \rightarrow A_2) \rightarrow B(A_2)$

(3) $\vdash B(A) \rightarrow B(B(A))$

(1) says that if the system has the power to assert A , it has the power to assert $B(A)$. (2) says essentially that the system knows it has the power to reason by modus ponens. (3) says, in effect, that the system knows (1). All of these assumptions seem unproblematic. To these we add the key assumption:

(4) $\vdash \text{not } B(\text{false})$

which says that the system asserts that it is not inconsistent. It turns out that these assumptions, along with the assumption that the system has the resources to do Peano arithmetic, lead to a contradiction.

To see this, we simply construct a sentence G such that

(5) $\vdash G \leftrightarrow \text{not } B(G)$.

This is a standard diagonal construction, and does not rely on any assumptions about the system's computability. We define the function *diag* in Peano arithmetic so that *diag*($\ulcorner C(x) \urcorner$) is $\ulcorner C(\ulcorner C(x) \urcorner) \urcorner$ for any predicate C . (For clarity, I reintroduce the $\ulcorner$ notation for Gödel numbering.) Then let G be the sentence *not* $B(\text{diag}(\ulcorner \text{not } B(\text{diag}(x)) \urcorner))$. It is straightforward to show that $G \leftrightarrow \text{not } B(\ulcorner G \urcorner)$. As long as the system

has at least the capacities of Peano arithmetic, it can replicate this reasoning, so that $\vdash G \leftrightarrow \text{not } B(\ulcorner G \urcorner)$.

G is effectively a sentence that says "I do not believe G ", much like a standard Gödelian construction, but without any assumptions about computability. It is not hard to see how the contradiction arises. The system knows that if it believes G , it is unsound; so it knows that if it is sound, it does not believe G . But this is to say that it knows that if it is sound, G is true. By assumption, it knows that it is sound, so it knows that G is true. So now it must be unsound, as it has fallen into a contradiction. This reasoning is easily formalized:

(6) $\vdash B(G) \rightarrow B(\text{not } B(G))$ [from (5), (1), (2)]

(7) $\vdash B(G) \rightarrow B(B(G))$ [from (3)]

(8) $\vdash B(G) \rightarrow B(\text{false})$ [from (6), (7), (2)]

(9) $\vdash B(\text{false}) \rightarrow B(G)$ [from (2), along with $\vdash B(\text{false} \rightarrow G)$]

(10) $\vdash G \leftrightarrow \text{not } B(\text{false})$ [from (5), (8), (9)]

(11) $\vdash B(G)$ [from (10), (4), (1)]

(12) $\vdash B(\text{false})$ [from (11), (9)]

We can see, then, that the assumption that we know we are sound leads to a contradiction. One might try to pin the blame on one of the other assumptions, but all these seem quite straightforward. Indeed, these include the sort of implicit assumptions that Penrose appeals to in his arguments all the time. Indeed, one could make the case that all of premises (1)-(4) are implicitly appealed to in Penrose's main argument. For the purposes of the argument against Penrose, it does not really matter which we blame for the contradiction, but I think it is fairly clear that it is the assumption that the system knows that it is sound that causes most of the damage. It is this assumption, then, that should be withdrawn.

Penrose has therefore pointed to a false culprit. When the contradiction is reached, he pins the blame on the assumption that our reasoning powers are captured by a formal system F . But the argument above shows that this assumption is inessential in reaching the contradiction: A similar contradiction, via a not dissimilar sort of argument, can be reached even in the absence of that assumption. It follows that the responsibility for the contradiction lies elsewhere than in the assumption of computability. It is the assumption about knowledge of soundness that should be withdrawn.

Still, Penrose's argument has succeeded in clarifying some issues. In a sense, it shows where the *deepest* flaw in Gödelian arguments lies. One might have thought that the deepest flaw lay in the unjustified claim that one can see the soundness of certain formal systems that underlie our own reasoning. But in fact, if the above analysis is correct, the deepest flaw lies in the assumption that we know that we are sound. All

Gödelian arguments appeal to this premise somewhere, but in fact the premise generates a contradiction. Perhaps we are sound, but we cannot know unassailably that we are sound.

3. The Missing Science of Consciousness?

A reader who is not convinced by Penrose's Gödelian arguments is left with little reason to accept his claims that physics is noncomputable and that quantum processes are essential to cognition, although these speculations are interesting in their own right. But even if one accepts that human behavior can be accounted for computationally, there is still the question of human consciousness, which after all is Penrose's ultimate target.

Penrose is clear that the puzzle of consciousness is one of his central motivations. Indeed, one reason for his skepticism about AI is that it is so hard to see how the mere enactment of a computation should give rise to an inner subjective life. Why couldn't all the computation go in the dark, without consciousness? So Penrose postulates that we to appeal to physics instead, and suggests that the locus of consciousness may be a quantum gravity process in microtubules. But this seems to suffer from exactly the same problem. Why should quantum processes in microtubules give rise to consciousness, any more than computational processes should? Neither suggestion seems appreciably better off than the other.

Although Penrose's quantum-gravity proposal might at least *conceivably* help explain certain elements of human behavior (if behavior turned out to be uncomputable, for example), it simply seems to be the wrong sort of thing to explain human consciousness. Indeed, Penrose nowhere claims that it does, and by the end of the book the "Missing Science of Consciousness" seems as far off as it ever was. As things stand, even by the end of Penrose's book, we seem to be left in Penrose's position *D*: these physical theories leave consciousness entirely unexplained.

This might seem odd, given that Penrose says he embraces position *C*, but in fact *C* and *D* are quite compatible. This is because Penrose's four positions run together a number of separate issues. For convenience, I repeat the positions here:

A: All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.

B: Awareness if a feature of the brain's physical action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness.

C: Appropriate physical action evokes awareness, but this physical action cannot even be properly simulated computationally.

D: Awareness cannot be explained by physical, computational, or any other scientific terms.

Note that *A*, *B*, and *C* all concern how awareness is *evoked*, but *D* concerns how awareness is *explained*.

These are two very different issues. To see the contrast, note that almost everybody would accept that the brain *evokes* awareness - if we were to construct a duplicate brain, there would be conscious experience associated with it. But it is far from clear that a physical description of the brain can *explain* awareness - many people have argued that given any physical account of brain processes, the question of *how* those processes evoke conscious experience will be unanswered by the physical account.

To really clarify the positions in the vicinity, we have to distinguish three questions:

- (1) What does it take to simulate our physical *action*?
- (2) What does it take to *evoke* conscious awareness?
- (3) What does it take to *explain* conscious awareness?

In answer to each question, one might say that (a) Computation alone is enough, (b) Physics is enough, but physical features beyond computation are required, or (c) Not even physics is enough. Call these positions C, P, and N. So we have a total of 27 positions, that one might label CCC, CPN, and so on.

Question (1) is the question Penrose is concerned with for most of the book, and the issue that separates *B* and *C* above. He argues for position P-- over C--. Descartes might have argued for N--, but few would embrace such a position these days.

Question (2) is the issue at the heart of Searle's Chinese room argument, and the issue that separates *A* from *B* and *C* above. Searle argues for -P- over -C-, and Penrose is clearly sympathetic with this position. Almost everyone would accept that a physical duplicate of me would "evoke" consciousness, so position -N- is not central here.

Question (3) is the central question about the *explanation* of consciousness (a question that much of my own work is concerned with). Penrose's positions *A*, *B*, and *C* are neutral on this question, but *D* is solely concerned with it; so in a sense, *D* is independent of the rest. Many advocates of AI might hold --C, some neurobiologists might hold --P, whereas my own position is --N.

The four positions Penrose describes come down to CC- (*A*), CP- (*B*), PP- (*C*), and --N (*D*). Penrose seems to think that in arguing for position *C* (PP-) he is arguing against position *D* (--N), but it is clear from this analysis that this is not so. In the end, nothing in Penrose's book bears on question (3), which is a pity, though it is certainly understandable. It would be very interesting to hear Penrose's position on just how physical theories might or might not explain human consciousness.

Indeed, one might even combine positions *A* and *D*, as I do, embracing CCN. On this position, human-like behavior can be produced computationally, and indeed enacting the right computation will give rise to consciousness, but neither a computational account nor a physical account alone will explain consciousness. It might seem odd that computation should evoke but not explain consciousness, but this

is no more odd than the corresponding position that neurophysiology might evoke but not explain consciousness. In either case, consciousness emerges from some underlying basis, but we need a further element in the theory to explain just how and why it emerges.

One can have a lot of fun cataloging positions (Dennett is CCC; Searle may be CPP; Eccles is NNN; Penrose is PPP; I am CCN; some philosophers and neuroscientists are CPN or PPN; note that all these are "non-decreasing" in $C \rightarrow P \rightarrow N$, as we might expect), but this is enough for now. The main point is that Penrose's treatment runs together question (3) with questions (1) and (2), so that in the end the question of how consciousness might be explained is left to one side.

A true science of consciousness will have address all of these questions, and especially question (3). Penrose has produced an enormously enjoyable and challenging book, but it seems to me that for all his hard work, the science of consciousness is still missing.

Miscellaneous Writings (David Chalmers).

This page contains a number of miscellaneous unpublished items from over the years.

The two-envelope paradox

- [The Two-Envelope Paradox: A Complete Analysis? \(PS\)](#)
 - You're given two envelopes, and are told that one contains twice the amount in the other, but you aren't told which is which. You tentatively decide to take envelope A, but then reason that there is a 50% chance that B contains twice A's amount and a 50% chance that it contains half A's amount, with an expected value of 1.25 times A's amount overall, so it is in your interests to switch. But of course the same holds in reverse. What is going on? I give a detailed analysis of this "two-envelope" paradox, including a few interesting subtleties that are sometimes overlooked. I didn't publish this as I now think the analysis is incomplete; in effect it solves the "numerical" paradox but not the "decision-theoretic" paradox.
- [The St. Petersburg Two-Envelope Paradox](#)
 - This is a much more recent short paper (2001) that fills in the gap in the paper above. It sets out a closely related scenario, which combines elements of the St. Petersburg paradox and the two-envelope paradox, and use this to diagnose where the decision-theoretic paradoxical reasoning goes wrong.

Old proto-papers

These are a few unfinished papers on a variety of subjects, mostly written around when I was graduate student at Indiana. I set these aside because I wasn't quite happy with them, or because I didn't think they were very significant, or because they evolved into something else, or because I am lazy, but looking at them now I quite enjoy them. My 1999 self may not endorse every youthful excess here, but there are some interesting bits and pieces.

- [The First-Person and Third-Person Views](#)
 - An early high-level overview of "first-person" and "third-person" issues about consciousness. This was Part I of a supposedly 3-part paper - the two remaining parts got turned into my paper "[Consciousness and Cognition](#)". This part doesn't reach any firm conclusions, but it captures something of the eternal internal struggle.
- [A Taxonomy of Cognitive Jokes](#)

- This was my inner taxonomist at work. (See my [bibliography](#) for another example of this sad trait.) Following a workshop on humor at CRCC, during which innumerable jokes were told and dissected, this was my attempt to fit all of the jokes into a few basic templates. Unfortunately the jokes themselves are given only one-line summaries, so I wish you luck with recognizing them.
- [On Spaghetti-Sorters and the Powers of Analog Computation](#)
 - This has a few jottings on the existence of "spaghetti-sorters" and other remarkable devices that seem to be able to compute certain functions faster than any standard computer (a spaghetti-sorter sorts a list of n numbers in order n time, for example, by cutting n pieces of spaghetti with the right lengths and banging them against a table).
- [How Cartesian Dualism Might Have Been True](#)
 - On a not-too-far-fetched scenario, inspired by contemporary work in artificial life, in which a Cartesian interactive dualism would have turned out to be true. Indeed, many of the "simulated worlds" in artificial life are Cartesian in just this way.

Philosophical commentaries

These are a few commentaries I've given at philosophy conferences over the years. Often I don't write these up, but these may have enough in them to be worthwhile.

- [Connectionist Representation and Deep Systematicity](#)
 - A 1991 commentary on Andy Clark's "Theoretical Spaces". I think the idea of "deep systematicity" as the central virtue of connectionism is quite important, although I haven't got around to developing it in a real paper. There are also some remarks on the relationship between evolution and learning.
- [Determining the Moment of Consciousness?](#)
 - This was a 1993 commentary on [Valerie Hardcastle](#)'s paper "On Determining the Moment of Consciousness" (recently published in [Philosophical Psychology](#)), which argues the neural level can help determine some "Orwell/Stalin" indeterminacies. Here I was in the unusual position of defending Dan Dennett; at the end Dennett was in the unusual position of agreeing with almost everything I said.
- [What is it Like to be a Thermostat?](#)
 - A 1994 commentary on Dan Lloyd's "What is it Like to be a Net?" Most people at the time thought I was joking.

Old Usenet postings

In days gone by I posted to Usenet, and other mailing lists, quite a bit. There was an occasional very good discussion - at least one of my published papers resulted from this sort of thing, and I can see other traces here and there. Other postings were just intellectual curiosity or frivolity. The endless discussions of Chinese rooms and consciousness were too open-ended to excerpt here, but here are a few more specific topics.

- ["Pick a number between zero and infinity"](#)
 - About the time I stood on a street corner and asked just that.
- [Is the Continuum Hypothesis true, false, or neither?](#)
 - I asked this question in sci.math and got quite a few interesting responses. This message (which is now part of the [sci.math FAQ](#)) is a summary and discussion. Lots of information on the Continuum Hypothesis can be found [here](#).
- [Thoughts on emergence](#)
 - An attempt at explicating the notion of "emergence" that is very popular in artificial life, complex systems theory, and other parts of contemporary science.
- [Does a rock implement every FSA?](#)
 - A comp.ai.philosophy discussion of Putnam's argument that every ordinary open system implements every finite automaton. A number of people contributed to this discussion, which has lots of interesting byways and raises plenty of deep questions about computation. My [paper](#) of the same title descended from this.
- [Realms of cognitive science](#)
 - Someone else posted about various "realms" such as those of matter, representation, and consciousness; here I take the conceit and run with it, multiplying the realms and discussing various possible relations among them.

Go to:

- [David Chalmers' home page](#)

The Two-Envelope Paradox: A Complete Analysis?

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721

chalmers@arizona.edu

[[Doug Hofstadter introduced me to the two-envelope paradox in 1988. This paper corresponds to more or less the position I came up with then. I wrote this up in 1994 after a couple of papers on the subject appeared in *Analysis*. I never published it, partly because it came to seem to me that this treatment resolves only part of the paradox: it resolves the "numerical" paradox but not the "decision-theoretic" paradox. For a more recent treatment of the decision-theoretic paradox, see [The St. Petersburg Two-Envelope Paradox](#).]]

A wealthy eccentric places two envelopes in front of you. She tells you that both envelopes contain money, and that one contains twice as much as the other, but she does not tell you which is which. You are allowed to choose one envelope, and to keep all the money you find inside.

This may seem innocuous, but it generates an apparent paradox. Say that you choose envelope 1, and it contains \$100. In evaluating your decision, you reason that there is a 50% chance that envelope 2 contains \$200, and a 50% chance that it contains \$50. In retrospect, you reason, you should have taken envelope 2, as its expected value is \$125. If your sponsor offered you the chance to change your decision now, it seems that you should do so. Now, this reasoning is independent of the actual amount in envelope 1, and in fact can be carried out in advance of opening the envelope; it follows that whatever envelope 1 contains, it would be better to choose envelope 2. But the situation with respect to the two envelopes is symmetrical, so the same reasoning tells you that whatever envelope 2 contains, you would do better to choose envelope 1. This seems contradictory. What has gone wrong?

The paradox can be expressed numerically. Let A and B be the amounts in envelope 1 and 2 respectively; their expected values are $E(A)$ and $E(B)$. For all n , it seems that $p(B > A | A = n) = 0.5$, so that $E(B | A = n) = 1.25n$. It follows that $E(B) = 1.25E(A)$, and therefore that $E(B) > E(A)$ if either expected value is greater than zero. The same reasoning shows that $E(A) > E(B)$, but the conjunction is impossible, and in any case $E(A) = E(B)$ by symmetry. Again, what has gone wrong?

This problem has been discussed in the pages of *Analysis* by Jackson, Menzies and Oppy [2], and by Castell and Batens [1], but for reasons that will become clear I think that their analyses are incomplete and mistaken respectively, although both contain insights that are important to the resolution of the problem. I will therefore present my own analysis of the "paradox" below.

Some distractions inessential to the problem arise from the facts that in the real world, money comes in discrete amounts (dollars and cents, pounds and pence) and that there are known limits on the world's money supply. We can remove these distractions by stipulating that for the purposes of the problem, the amounts in the envelopes can be any positive real number.

There are a number of steps in the resolution of the paradox. The first step is to note (as do the authors mentioned above) that the amounts in the envelopes do not fall out of the sky, but must be drawn from some probability distribution. Let the relevant probability density function be g , where the probability that the smaller amount falls between a and b is $\int_a^b g(x) dx$. We can think of this distribution as either representing the chooser's prior expectations, or as the distribution from which the actual values are drawn. I will generally write as if it is the second, but nothing much rests on this. To fix ideas, we can imagine that our sponsor chooses a random variable Z with probability density g , and then flips a coin. If the coin comes up heads, she sets $A=Z$ and $B=2Z$; if it comes up tails, she sets $A=2Z$ and $B=Z$.

Recognizing the existence of a distribution immediately shows us that the reasoning that leads to the paradox is not always valid, as Jackson *et al* note. For example, if the distribution is a uniform distribution over values between 0 and 1000, with amounts over 1000 being impossible, then if $A > 500$, it is always a bad idea to switch. It is therefore not true that for all distributions and all values of n , $p(B>A|A=n) = 0.5$. In general, $E(B|A=n)$ will not depend only on n ; it will also depend on the underlying distribution.

In their analysis, Jackson *et al* are satisfied with this observation, combined with the observation that limitations on the worlds' money supply ensure that in practice the relevant distributions will always be bounded above and below. The paradox does not arise for bounded distributions, as we saw above. When A is a medium value, there may be equal chances that B is larger or smaller, but when A is large B is likely to be smaller, and when A is small B is likely to be larger, so the paradox does not get off the ground.

This practical observation is an insufficient response to the mathematical paradox, however, as Castell and Batens note. Unbounded distributions can exist in principle if not in practice, and in-principle existence is all that is needed for the paradox to have its bite. For example, it might seem that if the distribution were a uniform distribution over the real numbers, then $p(B>A|A=n) = 0.5$ for all n . This would seem to have paradoxical consequences for mathematics, if not for the world's money supply.

This leads to the second step in the resolution of the paradox, which is that taken by Castell and Batens. (We will see that this step is ultimately inessential to the paradox's resolution, but it is an important intermediate point of enlightenment.) There is in fact no such thing as a uniform probability distribution over the real numbers. To see this, let g be a uniform function over the real numbers. Then $\int_k^{k+1} g(x)dx$ is equal to some constant c for all k . If $c=0$, then the area under the entire curve will be zero, and if $c>0$, then the area under the entire curve will be infinite, both of which contradict the requirement that the integral of a probability distribution be 1. At one point Jackson *et al* raise the possibility of infinitesimal probabilities, but if this is interpreted as allowing c to be infinitesimal, the

suggestion does not work any better. To see this, note that if the distribution is uniform:

$$\begin{aligned} \int_0^{\infty} g(x) dx &= \int_0^1 g(x) dx + \int_1^2 g(x) dx + \int_2^3 g(x) dx + \dots \\ &= \int_0^1 g(x) dx + \int_2^3 g(x) dx + \int_4^5 g(x) dx \\ &= (\int_0^{\infty} g(x) dx)/2 \end{aligned}$$

so that the overall integral must be zero or infinite. A uniform distribution over the real numbers can only be an "improper" distribution, whose overall integral is not 1.

The impossibility of a uniform probability distribution over the real numbers is reflected in the fact that every proper distribution must eventually "taper off": for all $\epsilon > 0$, there must exist k such that $\int_k^{\infty} g(x) dx < \epsilon$. It is very tempting to suppose that this "tapering off" supplies the resolution to the paradox, as it seems to imply that if A is near the high end of the (proper) distribution, it will be more likely that B is smaller; perhaps sufficiently more likely to offset the paradoxical reasoning? This is the conclusion that Castell and Batens draw. They offer a "proof" that the distribution must be improper for the paradoxical reasoning to be possible.

Unfortunately Castell and Batens' proof is mistaken, and in fact there exist proper distributions for which the paradoxical reasoning is possible. The error lies in their assumption, early in the paper, that $p(B > A | A = n) = g(n)/(g(n) + g(n/2))$. This seems intuitively reasonable, but in fact $p(B > A | A = n) = 2g(n)/(2g(n) + g(n/2))$, which is significantly larger in general.

To see this, note that if A is in the range $n \pm dx$, then B is either in the range $2n \pm 2dx$ or in the range $n/2 \pm dx/2$. The probability of the first, relative to the initial distribution, is $g(n)dx$; the probability of the second is $g(n/2)dx/2$. The probabilities that B is greater or less than A therefore stand in the ratio $2g(n):g(n/2)$, not $g(n):g(n/2)$, as Castell and Batens suppose.

For example, given a uniform distribution between 0 and 1000, if A is around 100, it is in fact twice as likely that B is around 200 than that B is around 50. To dispel any lingering counterintuitiveness, note that something like this *has* to be the case to make up for the fact that when $A > 500$, B is always less than A . To find a distribution where the chances of a gain and a loss are truly equal for many n , we should turn not to a uniform distribution but to a decreasing distribution, where $g(n/2) = 2g(n)$ for many n . An example is the distribution $g(x) = 1/x$, where we cut off the distribution between arbitrary bounds L and U , and normalize so that it has an integral of 1. This distribution will have the property that for all n such that $2L < n < U/2$, $p(B > A | A = n) = 0.5$. To illustrate this intuitively, note that for such a decreasing distribution, the prior probability that the smaller value is between 4 and 8 is the same as the probability that it is between 8 and 16, and so on, if L and U are appropriate. Given the information that $8 < A < 16$, it is equally likely that B is in the range above or below.

This flaw in Castell and Batens' reasoning nullifies their proof that a distribution must be improper for the paradoxical reasoning to arise, but it does not yet show that the conclusion is false. It remains open whether there is a proper distribution for which the paradoxical reasoning is possible. The bounded

distribution above will not work, as its bound will block the paradoxical reasoning in the usual fashion; and the *unbounded* distribution $g(x) = 1/x$ is improper, having an infinite integral. But this can easily be fixed, by allowing the distribution to taper off slightly faster. In particular, the distribution $g(x) = x^{-(1.5)}$, cut off below a lower bound L and normalized, allows the paradox to arise. The distribution has a finite integral, and even though for most n , $p(B > A | A = n) < 0.5$, it is still the case that for all relevant n , $E(B | A = n) > n$. To see this, note that if $n < 2L$, then $E(B | A = n) = 2n$; and if $n \geq 2L$, then

$$\begin{aligned} p(B > A | A = n) : p(B < A | A = n) &= 2g(n) : g(n/2) \\ &= 2n^{-(1.5)} : (n/2)^{-(1.5)} \\ &= 1 : \sqrt{2}. \end{aligned}$$

The expected value $E(B | A = n)$ is $(2n + \sqrt{2}n/2)/(1 + \sqrt{2})$, which is about $1.12n$. The paradox therefore still arises.

The distribution here may be unintuitive, but it is easy to illustrate a similar distribution intuitively. Take a distribution in which the probability of a value between 1 and 2 is c , the probability of a value between 2 and 4 is just slightly less, say $0.9c$, the probability of a value between 4 and 8 is $0.81c$, and so on. This distribution has a finite integral, as the integral is the sum of a decreasing geometric series; and it is sufficiently close to the case in which the probability of a value between 2^k and 2^{k+1} is constant that the paradoxical reasoning still arises. Even though $p(B < A | A = n)$ is now slightly less than 0.5, due to the incorporated factor of 0.9, it has decreased by a sufficiently small amount that $E(B | A = n)$ remains greater than n . The case $g(x) = x^{-(1.5)}$ is just like this, except that the factor of 0.9 is replaced by a factor of $1/\sqrt{2}$, which is around 0.7.

The paradox has therefore not yet been vanquished; there are perfectly proper distributions for which the paradoxical reasoning still applies. This leads us to the third and final step in the resolution of the paradox. Note that although the distributions above have finite integrals, as a probability distribution should, they have infinite *expected value*. The expected value of a distribution is $\int_0^{\infty} xg(x)dx$. When $g(x) = x^{-(1.5)}$ (cut off below L), the expected value is $\int_L^{\infty} x^{-(0.5)} dx$, which is infinite. But if the expected value of the distribution is infinite, there is no paradox! There is no contradiction between the facts that $E(B) = 1.12 E(A)$ and $E(A) = 1.12 E(B)$ if both $E(A)$ and $E(B)$ are infinite. Rather, we have just another example of a familiar phenomenon, the strange behavior of infinity.[*]

*[[[Castell and Batens note some similar consequences of infinite expected values in another context, in which the distribution is over a countable set. They say that infinite expected values are "absurd", but I do not see any mathematical absurdity.]]]

To fully resolve the paradox, we need only demonstrate that for distributions with finite expected value, the paradoxical situation does not arise. To do this, we need to precisely state the conditions expressing the paradoxical situation. In its strongest form, the paradoxical situation arises when $E(B | A = n) > n$ for all n . However, it arises more generally whenever reasoning from B 's dependence on A leads us to the

conclusion that there is expected gain *on average* (rather than all the time) by switching A for B . This will hold whenever $E(K-A) > 0$, where K is the random variable derived from A by the transformation $x \rightarrow E(B|A=x)$. We therefore need to show that when $E(A)$ is finite, $E(K-A) = 0$.

Let h be the density function of A . Then $h(x) = (g(x) + g(x/2))/2 = (2g(x) + g(x/2))/4$. (Note that $h \neq g$, as g is the density function of the *smaller* value.) Then

$$\begin{aligned} E(K-A) &= \int_0^{\infty} h(x) (E(B|A=x) - x) dx \\ &= \int_0^{\infty} (2g(x) + g(x/2))/4 \cdot ((2x \cdot 2g(x) + x/2 \cdot g(x/2))/(2g(x) + g(x/2)) - x) dx \\ &= \int_0^{\infty} (2xg(x) - x/2 \cdot g(x/2))/4 dx \\ &= (\int_0^{\infty} 2xg(x) dx - \int_0^{\infty} 2yg(y) dy)/4 \\ &= 0. \end{aligned}$$

Note that the fourth and fifth steps above are valid only if $\int_0^{\infty} xg(x) dx$ is finite, which holds iff $E(A)$ is finite. (If $\int_0^{\infty} xg(x) dx$ is infinite, it is possible that $\int_0^{\infty} 2xg(x) - x/2 \cdot g(x/2) dx \neq 0$, even though $\int_0^{\infty} 2xg(x) dx = \int_0^{\infty} x/2 \cdot g(x/2) dx$.)

It follows that when $E(A)$ is finite, consideration of the dependence of B on A will not lead one to the conclusion that one should switch A for B . A colollary of the result is that when $E(A)$ is finite, it is impossible that $E(B|A=n) > n$ for all n , so that the strong form of the paradox certainly cannot arise.

If $E(A)$ is infinite, this result does not hold. In such a case, it is possible that $E(A) = E(K)$ (both are infinite) but that $E(K-A) > 0$. Here, the "paradoxical" reasoning will indeed arise. But now the result is no longer paradoxical; it is merely counterintuitive. It is a consequence of the fact that given infinite expectations, *any* given finite value will be disappointing. The situation here is somewhat reminiscent of the classical St. Petersburg paradox: both "paradoxes" exploit random variables whose values are always finite, but whose expected values are infinite. The combination of finite values with infinite expected values leads to counterintuitive consequences, but we cannot expect intuitive results where infinity is concerned.[*]

References

- [1] P. Castell and D. Batens, 'The Two-Envelope Paradox: The Infinite Case'. *Analysis* 54:46-49.
- [2] F. Jackson, P. Menzies, and G. Oppy, 'The Two Envelope "Paradox"', *Analysis* 54:43-45.

The St. Petersburg Two-Envelope Paradox

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

The Two-Envelope Paradox: I am presented with two envelopes A and B, and told that one contains twice as much money as the other. I am given envelope A, and offered the options of keeping envelope A or switching to B. What should I do?

I reason: (1) For any x , if I knew that A contained x , then the odds are even that B contains either $2x$ or $x/2$, so the expected amount in B would be $5x/4$. So (2) for all x , if I knew that A contained x , I would have an expected gain in switching to B. So (3) I should switch to B. But this seems clearly wrong, as my information about A and B is symmetrical.

It is widely accepted that one's reasoning in this scenario should depend on one's initial probability distribution for the amounts in the envelopes. It is also widely accepted that if this distribution has a finite expected value, (1) and (2) are false, so the reasoning goes wrong. However, it is well-known that there are certain distributions with infinite expected value such that although (1) is false, (2) is still true. For such distributions, the paradox seems to arise as strongly as ever, and there is no widespread agreement on how to handle it.[*]

*[[For the existence of paradoxical distributions with infinite expected values, see Broome 1994. Chalmers 1994 gives a proof that the paradox does not arise for finite distributions, and notes the existence of infinite paradoxical distributions. I now think that the discussion there does not get to the bottom of the residual "paradox"; the current paper is a corrective.]]

To help diagnose the residual "paradox", I offer the following variant, obtained by mixing and matching two well-known scenarios:

The St. Petersburg Two-Envelope Paradox: I am presented with two envelopes, A and B. I am told that each of them contains an amount determined by the following procedure, performed separately for each envelope: a coin was flipped until it came up heads, and if it came up heads on the n th trial, 2^n is put into the envelope. This procedure is performed separately for each envelope. I am given envelope A, and offered the options of keeping A or switching to B. What should I do?

I reason: (0) Before opening the envelopes, the expected value in each is infinite. (1) For any x , if I knew that A contained x , then the expected value in B would still be infinite. So (2) for all x , if I knew that A

contained x , I would have an expected gain in switching to B. So (3) I should switch to B. But this seems clearly wrong, as my information about A and B is symmetrical.

What is going on here? Obviously the problem lies in the step from (2) to (3). When a distribution over finite amounts has an infinite expected value, any specific result will be disappointing. It will then always be in one's interest to do things over, if given the opportunity. But it does not follow from this that *before* knowing the result, it is in one's interest to do things over. The general moral can be put decision-theoretically or probabilistically.

Decision-theoretically, the moral is that dominance reasoning is not generally valid. Say one has two choices A and B, and a parameter C whose value is unknown, on which the utility of A or B may depend. An unrestricted dominance principle says that if, given any specific value of C, A is preferable to B, then A is preferable to B overall. The St. Petersburg two-envelope paradox (taking C as the value in envelope A, and A and B as the respective choices) shows us that this principle is false.

Probabilistically, the moral is that one cannot discharge certain universally quantified claims about conditional expected values. An initially plausible principle holds that given any two unknown quantities C and D, then if $E(D|C=x) > 0$ for all x , then $E(D) > 0$. The St. Petersburg two-envelope paradox (taking $C = A$, and $D = B-A$, where A and B are the amounts in the two envelopes, and allowing infinite positive and negative expected values) shows that this principle is false.

What applies to the St. Petersburg two-envelope paradox applies equally to the original two-envelope paradox. In both cases, the reasoning from (2) to (3) tacitly relies on a principle such as those above. These principles are false, so the reasoning is invalid.

Why do the principles fail in these cases? In both cases, the overall expected gain from switching (B-A) can be represented as the sum of an infinite series in which the sums of both the positive and the negative terms diverge. Such a series has an undefined sum, but can always be partitioned into a new series with only positive terms (each of which is a sum of terms in the original series) and also into a new series with only negative terms. (In the envelope cases, one obtains such new series via partitions that hold fixed the value of A and the value of B respectively.) When this instability under partitioning is present, it is to be expected that dominance reasoning and the principles above break down.

These principles may be true under certain restrictions: e.g., a restriction to cases involving only finite expected values and absolutely convergent series (setting aside Newcomb-style cases, in which C is not independent of one's decision between A and B). There may be weaker restrictions that suffice, since there are plausibly some cases involving infinite expected values where dominance reasoning is valid. Articulating the weakest possible restrictions here is an important open question; but for now, what matters is that the principles do not hold in general.

The upshot is a disjunctive diagnosis of the two-envelope paradox. The expected value of the amount in the envelopes is either finite or infinite. If it is finite, then (1) and (2) are false: the paradoxical reasoning

results from illicitly ignoring probability distributions. If it is infinite, then the step from (2) to (3) is invalid: the paradoxical reasoning relies on dominance principles or probabilistic principles that are false in the infinite case.[*]

*[[Apart from the illustration provided by the St. Petersburg two-envelope paradox, most of the issues discussed here have been discussed elsewhere: Dreier (forthcoming) discusses the breakdown of dominance reasoning; Norton (1998) discusses matters in the vicinity of the false probabilistic principle, and Clark and Shackel (2000) discuss the role of absolute convergence. Thanks to Michael Clark and Nicholas Shackel for comments and to Jamie Dreier for discussion.]]

References

Broome, J. 1995. The 2-Envelope Paradox. *Analysis* 55:6-11.

Chalmers, D.J. 1994. The two-envelope paradox: A complete analysis? [consc.net/papers/envelope.html]

Clark, M. & Shackel, N. 2000. The two-envelope paradox. *Mind* 109:415-42.

Dreier, J. (forthcoming). Boundless good.

Norton, J.D. 1998. When the sum of our expectations fails us: The exchange paradox. *Pacific Philosophical Quarterly* 79:34-58.

The First-Person and Third- Person Views (Part I)

David J. Chalmers

[[An overview of "first-person" and "third-person" issues about consciousness, written when I was a first-year graduate student at Indiana in 1989. This was Part I of a supposedly 3-part paper - the two remaining parts got turned into "[Consciousness and Cognition](#)". This part doesn't reach any firm conclusions, but it captures something of the eternal internal struggle.]]

Organization, roughly.

- Intro to what "first-person" and "third-person" mean.
 - (outline the probs of the first person)
 - (convenience of third-person vs absoluteness of first-person)
 - (explain terminology)
- Dominance of third-person, reasons.
 - (embarrassment with first person)
 - (division of reactions)
 - (natural selection - those who can make the most noise)
 - (analogy with behaviourism)
- Reductionism, hard-line and soft-line
- Appropriation of first-person terms by reductionists.
 - (mind, consciousness, first-person)
 - (correspondence of first-person to third-person)
- The problems of the first person
 - (problem of qualia, problem of mental content)
- The cult of the first person
- The reconciliation - what the third person can tell us about the first person

We CAN'T separate the first-person from the third-person. Even as I write this, it's my brain that is doing the thinking.

The First Person and the Third Person

Perhaps the most important duality in the philosophy of mind is that between the first-person and third-person views of mental events. Some might say that the fundamental duality is that between mind and brain, or between subjective and objective - but all of these reduce to the first-person/third-person duality.

The first person is, at least to many of us, still a huge mystery. The famous "Mind-Body Problem," in these enlightened materialist days, reduces to nothing but the question "What is the first person, and how is it possible?". There are many aspects to the first-person mystery. The first-person view of the mental

encompasses phenomena which seem to resist any explanation from the third person. Such phenomena include some famous philosophical bugbears: subjective experience, qualia, consciousness, and even mental content (although see below. It is notoriously difficult to even talk about the first person without slipping into confusion, and these terms may have different connotations for different people.).

The distinguishing mark of the first-person view is the air of mystery which surrounds it. This feeling of mysteriousness has led many people to dismiss the first-person out of hand. It perhaps has 'spiritual' connotations not unlike those of the occult or religion. But the first-person is not to be dismissed so easily. It is indeed a glaring anomaly today, in the heyday of the scientific world-view. If it was not for the direct *experience* which all of us have of the first-person, it would seem a ridiculous concept. But it throws up too many problems to be neatly packaged away in the kind of third-person explanation which suffices for everything else in the scientific world. Pity.

The third-person view, by contrast, poses no deep metaphysical difficulties. The difficulties here, while not to be underestimated, are in a sense merely technical. From the third-person view, the human brain is (in principle at least) perfectly understandable. It is after all only a physical system - a tremendously complex one, without doubt, but a physical system nevertheless - which like all other such systems, is constituted at the bottom line by microscopic physical parts, obeying the laws of physics. Its behaviour would be totally analyzable and predictable if one had a detailed knowledge of physical principles. From the third person, there is no room for any mystery.

Of course, as with any complex system, a microscopic-level physical explanation may not tell the whole story, or at any rate it may not tell the story as elegantly and simply as it could be told. Purely reductionist explanations are not noted for their parsimony. So third-person theorizers are drawn into a search for the correct level of explanation, and for useful high-level predicates and properties which can be abstracted from the physical base. Some kinds of abstractions may prove to have great explanatory and predictive power, in which case they will be retained; others will not, and will thus be discarded. This is the way of the world in any sort of complex system analysis - whether it be in the field of biology, economics or psychology.

The kind of abstractions which have proved particularly interesting to third- -person theorizers in the philosophy of mind include such diverse concepts as beliefs and desires, rationality, understanding,...; and occasionally even concepts like consciousness, sensation and experience (but importantly, these latter concepts have a subtly different emphasis when viewed from the third person to when viewed from the first person). Some of these have proved to have great explanatory power. But nevertheless, from the third person these remain (merely) *convenient* and *useful* abstractions. From the third person, no-one need ascribe any absolute 'reality' to these concepts, and more than is ascribed to any other abstraction. One might say that the distinguishing mark of the third-person view of the mental is *convenience*.

By contrast, first-person mental events seem to be something more than just convenient abstractions. When we talk of colour-sensations, or pains, or even consciousness, we are not doing this because of the need to make useful abstractions. We are doing this because they seem to have an absolute *reality*. From a totally 'objective' view, no-one would stop for a moment to think that there might really be something in

there *experiencing* those convenient abstractions which they had made; no third-person theorizer would believe that their elegant *stipulations* of mental content had any absolute truth behind them; nobody would believe that a mere physical system could really carry *meaning*; and yet there is, they do and it does. What would seem ridiculous from the objective view is almost self-evident from the subjective view. The first person is an anomaly, but it is forced upon us.

In recent years there has been a polarity between philosophical thinkers who take the third-person stance, and those who address the first-person issues. In a sense there seems to be an irreconcilable division. I will argue that this division is not as great as it seems to be. Unlike hard-line members of the third-person camp, I believe that the first-person does present us with major problems to be answered. Unlike hard-line members of the first- -person camp, I believe that the third-person view can be the source of much wisdom. The third-person and first-person are inseparable - I try to outline why, and how.

In Part I of this paper, I outline some of the major differences between the first-person and third-person camps, and some of the subdivisions within these camps. In some ways the polarity between these camps is not surprising - I try to make a case for why.

In Part II, I take up the third-person position. I give arguments showing why it was perhaps ridiculous to ever suppose that the first-person could be separated from the third-person view. In particular, I try to investigate the possibility of a being which *seems* to have all the third-person attributes required for, say, consciousness, but which is nevertheless lacking a first- -person 'experiencer.' More than one recent thinker seems to believe in such a possibility, but I show that if this could be the case, it would lead to strong consequences. In particular, it might mean that the Mind-Body Problem was insoluble.

By Part III, we will have seen the necessity for coherence between a first- -person and a third-person theory of mind. I will try to present a framework for such a theory, which might explain the notion of the first person without dismissing it. This will involve the two fundamental (and inseparable) concepts of *pattern* and *information*. Without giving too much away, the fundamental thesis will be this: third-person is to first-person as pattern is to information.

A Note on Terminology

Readers of this paper may hear rather more of the terms "first-person" and "third-person" than they would like. I originally intended to use the word "consciousness" to represent the mysteries of the first-person, but this had two problems: the word has been often used to denote third-person viewable phenomena (the notion of a system which gets feedback from its own processing, to name just one aspect), and this could lead to much confusion; and secondly, I believe there is more to the first-person mystery than consciousness alone (the problem of qualia, for instance, does not fit easily into a framework shaped around consciousness).

Instead I have chosen to use the term "first-person", as a coverall for all the problems which we recognize when we take the first-person stance - that range of problems often subsumed under the somewhat

antiquated heading "Mind-Body Problem." The baffling problems of qualia, consciousness, and subjective experience are all part of the "first-person" mystery. What this term lacks in sparkle, it makes up in generality and accuracy.

Many mental events are viewable from the first person and from the third person. Take a pain, for instance: from the first-person we are interested in the particular qualitative sensation to which it gives rise; from the third-person, we are interested in the state of the brain at the time (including the famous firing of C-fibres), its causal effect on other parts of the brain, and the behaviour to which it gives rise. Obviously, these views of a pain are very much related. But when I want to make the distinction between these aspects of a pain, I will refer to "first-person pains" and "third-person pains," as a kind of shorthand. Similarly for other mental events. Even such a phenomenon as consciousness we can view in two ways, as "first-person consciousness" and "third-person consciousness."

This is a very useful distinction to make. Firstly, it avoids confusion. Secondly, it allows us to talk about mental events from the third-person without *presupposing* that there are true, first-person mental events 'going on inside.' This is very useful when, for instance, we want to speculate on the possibility of a robot being conscious. The first step is to take what Dennett would call the "intentional stance" towards it, and see what kind of third-person mental events we could *ascribe*. Hard-line members of the first-person camp, such as Searle and McGinn, would deny us even this.

And an apology to grammaticists. "First-person" is an adjective, not a noun. When I speak of "the first person," this is usually a shorthand for "the first-person view," or sometimes "phenomena viewable from the first-person view." I hope the intended usage is clear in context. I do not mean to beg the question of whether there is indeed a single, indivisible "first person." This question, the problem of the self, is a related but separate issue which has been discussed by thinkers from Hume to Dennett and Parfit, and I believe the jury is still out.

The Dominance of the Third-Person View

The third-person view has enjoyed a distinct primacy in the philosophy of mind of recent years. This is not least because philosophers have finally caught up with the scientific notion that the bottom line of all that exists is the physical universe. But there are other reasons. In this section I will present some of the reasons why we should expect the third-person program to be dominant and quite successful. None of these reasons are inconsistent with the view that there remains a great mystery about the first person.

It seems that many philosophers of mind find first-person questions a little embarrassing. Here we have a very sophisticated academic field, but still it seems we have trouble getting far with what most outsiders would see as our fundamental problem: the Mind-Body Problem. There are four typical reactions to this embarrassment:

- (1) Ignore the first-person, and maybe hope that it will go away. If one is candid, admit one's embarrassment but lack of a solution. Work on other problems instead.

(2) Deny that there is any embarrassment at all. Claim that the mystery of the first-person was only ever a powerful illusion, perhaps a remnant of our Cartesian heritage. Attempt to demonstrate that a reductionistic (or eliminative) third-person analysis is enough to deal with the problems of the first-person. (This position is the Hard-Line Reductionist stance which I delineate below.)

(3) Admit the embarrassment, and spend one's time going around pointing out the problems. The only trouble with this approach is that without solutions to these problems, a new kind of embarrassment of one's own takes over. One can only point out the problems so many times without repeating oneself. The leading exponent of this stance has been Nagel. Searle's arguments have been a variation on this theme.

(4) Admit the embarrassment, and try to deal with the mystery of the first-person without denying its existence. This has been tried by many ambitious philosophers over the years, but no approach has met with universal approval. Many people recognize the *problems* of the first-person, but few seem satisfied with *solutions* that have been proposed. This makes attempting to resolve the first-person mysteries a somewhat quixotic enterprise. Consider this paper another addition to the canon.

Position (1) is the most common reaction. There is, after all, much interesting philosophy which can be done from the third-person stance. The sophistication of this domain contrasts hugely with the vagueness which surrounds the first-person. So we have today a proliferation of papers which deal with the third-person without taking a stance on the first person. Occasionally, authors explicitly admit to being baffled by the mysteries of the first person, but more often their position is left unstated. Some third-person philosophers par excellence include Fodor, Stich, ... (???).

Position (2) is also not too uncommon, and is a not unexpected reaction to the apparent mystery of the first person. There exist many recent papers which take the third-person stance, give a detailed analysis of mental phenomena, and in doing so claim to have dismissed or explained seemingly first-person problems. This claim is usually the part most in need of examination. Sometimes it is supported with a great deal of careful argument (Parfit). More often, what is given is a purely third-person account, perhaps an account of the architecture or evolution of mental structures (Dennett, Churchland); and immediately thereafter the *claim* is baldly made that this account eliminates the first-person mysteries. Often, at least to the eyes of a member of the first-person camp, the argument given to support this claim seems a little thin.

One of the main difficulties with taking position (3) or (4) is that it is difficult to even *talk* about the first-person without confusion. Part of the reason for this is the huge intrinsic difficulty of the problem; another reason is the predominance of terminological confusion when talking about such things as "mind," "consciousness," "self" and other traditionally first-person concepts. Everybody uses these terms in their own way, and there is no general agreement about their domains of applicability (let alone about to what they refer!). I go into more detail about this problem below.

Given this difficulty with talking about first-person issues, it is not surprising that third-person philosophy

of mind should be prevailing, almost by default. Despite the fact that many people out there are hugely puzzled by the problems of the first-person (as they may admit off-the-record), it is difficult to make much noise about it without falling into both repetition and negativity. The third-person stance, by contrast, can yield many important insights about the mind, despite the fact that it does not usually deal with the Big Issues. Taking the third-person stance, one is able to make much more noise. So, almost by natural selection, the third-person stance becomes dominant, and the first-person stance is reduced to the status of a quirky domain for those who like to cause trouble.

An Analogy with Behaviourism

One might make an analogy with behaviourism in psychology, in the earlier part of this century. In those days, nobody had the slightest idea what was going on inside the brain. To speak of internal cognitive processes was to descend to the level of speculation. The only things which people could get a handle on were behavioural manifestations. It was natural that as a result, behaviour became by far the dominant area of research. Nobody even had a clear idea of what would be the correct *language* to talk about internal processes with.

One may presume that when the behavioural paradigm was just beginning, not many serious thinkers actually doubted the *existence* of internal processes. It was just that nobody knew quite how to study them, and that in even talking about them researchers felt like they were standing on quicksand, descending into vagueness. Little wonder they grasped at whatever surer ground they had available, the direct study of behaviour. And indeed the analysis of behaviour produced many important insights. These insights had a profound and often positive effect on psychology. All that believers in 'cognitive processes' could do was stand around and say "but this isn't all!".

A curious thing began to happen. Where once behaviourists had abjured the study of internal processes because of the *difficulty* of dealing with these, some now, following upon the great success of their research program, began to deny the *existence* of internal processes at all. When spelt out, the argument of these radical behaviourists went something like this:

- (1) We have had much success analyzing behaviour in terms of stimulus- -response mechanisms, without needing 'internal processes.'
- (2) Believers in 'internal processes' have had little success in reaching any concrete conclusions, and seem vague about what they are even talking about.
- (3) There may be a few problems with our view, such as certain arguments proposed by the 'cognitivists,' but these will go away with careful analysis.

Therefore

- (4) 'Internal processes' are an unnecessary hypothesis, and do not exist.

Looking upon this with the assistance of hindsight, there seems to be a monumental non sequitur there somewhere. To be sure, behaviourism had met with some success, but surely no-one had ever expected it to tell us *everything*? And did they completely doubt the evidence of their own introspection, about the existence of internal states? The arguments of the cognitivists, which seem so obvious to us today, surely must have given them cause for doubt?

But we should not be surprised. The argument was almost self-supporting, by natural selection. Believers in internal processes found it difficult to get beyond naysaying, to achieve concrete results, because these processes were still shrouded in mystery. So, to make a living these believers had to either join the behaviourist bandwagon, or more likely leave the field. The behaviourists were the ones who were making the noise, given the constraints of the current state of scientific knowledge, so the field acquired an unnatural bias in their direction. To be sure, many 'reasonable' behaviourists remained, who admitted that one day internal processes would have to be dealt with; but these comments were more often made in the protection of their homes, or their departmental tea-room. The point could only be made in print so many times, before repetition and negativity set in.

Sound familiar? I am sure that I do not need to spell out the analogy. These days, nobody has any problem believing in cognitive processes. They may not be totally understood, but progress has been made and a language for discussion has been developed. The bugbear these days is "subjective experience," and related first-person phenomena. We are simply not yet sophisticated enough to discuss these without encountering a fog of vagueness. As with the behaviourists, this leads to a understandable predominance of third-person analysis. And just as with the behaviourists, there are some 'radicals' who, to avoid the embarrassment of problems we do not understand, claim that the problems do not exist, despite strong arguments to the contrary. We should not be surprised by this.

Reductionism and the cult of the third-person

The term "reductionism" is often a little vague in its application, especially in the philosophy of mind. Dawkins has said that "nobody is a reductionist in any sense worth being against." This may or may not be true in the philosophy of mind. I will divide reductionism into "hard-line" and "soft-line" varieties.

The Hard-Line Reductionist believes that to explain everything that is interesting about the mind, all we have to do is explain what is happening in the physical system that is the brain (and, perhaps, in the body and surrounding environment). Beyond this collection of third-person facts (they believe) there is *nothing more to explain*. Any claims of "but you've left out the most interesting part!" are not countenanced. These reductionists may allow certain abstractions from the physical base to be made, for the purposes of elegance and explanatory power, but these abstractions are made simply out of convenience, and they are not held to reflect any absolute truth.

Hard-line reductionists believe that questions of "subjective experience" and "qualia" and "phenomenology" are distracting chimeras, at best powerful illusions. A list of hard-line reductionists would include Ryle, Armstrong, and more recently, Dennett, Churchland and Parfit. These are people

who take position (2) above, in reaction to first-person issues: they deny that there is any mystery. I believe that this is a sense of reductionism is worth being against.

In contrast, Soft-Line Reductionists are much more amenable to first-person questions. The soft-line reductionists believes that the *cause* of everything which is going on in the mind is the physical system, and this physical system is explainable from the third-person, but there may still be some emergent phenomena which are not captured by a purely physical, third-person description. Most hard-line reductionists probably regard soft-line reductionists as unbearably "wimpy," in rather the same way that an atheist would regard an agnostic, or that a communist would regard a social democrat.

Essentially, a soft-line reductionist is a materialist who nevertheless believes that the first-person is a great mystery; a hard-line reductionist believes there is no mystery at all. While the soft-liner may be dismissed by the hard-liner as "soft, squishy, and mystical," in this paper I intend to show that not only is soft-line reductionism a tenable position, but also that it does not inevitably lead to throwing up ones armss in frustrated wonder at the mystery of it all. I believe that a theory of the first-person can be developed which is *coherent* with a theory of the third- -person, but which is not *subsumed* by a theory of the third-person.

The Appropriation of First-Person Terms

As I have said, the first-person is hard to talk about. One important reason is that every first-person concept has a *corresponding* third-person concept. The word "consciousness" has often been taken as a compact reference to all that is mysterious about the first-person; but obviously, there are some third-person aspects which are very relevant to the word. If consciousness means something like "awareness of self," then a third-person commentator can point to the properties of physical systems (like the brain) which are monitoring their own processing, and using this feedback to adjust their behaviour. "So," the reductionist will claim, "wasn't this just what you meant by the term?". The first-personite will of course reply "No, the problem is, how could a mere physical system *experience* this awareness." The reductionist will reply in character, and the two will go on, feeling that they are talking past each other. The two are talking about corresponding phenomena, but not necessarily about identical phenomena.

This direct correspondence between first-person and third-person phenomena is the cause of much of the confusion. It has led to a slippery terminological slope, where nobody is sure what words in the "mental" vocabulary are referring to at a given time. Increasingly, terms which were once reserved for first-person issues are now used to cover third-person issues also.

The word "mind" once stood for everything that was quintessentially first-person. Witness the phrase "mind-body problem," for instance. But over the years the emphasis of the term has changed, until now is now much more frequently used to refer to third-person phenomena. Terms in common parlance such as "the subconscious mind" bear witness to this fact. Cognitive science, which is essentially the third-person investigation of mechanisms of thought, is often described as the "study of mind." These days the word "mind" is a general coverall for abstractions from the brain, first-person or third-person.

The word "consciousness," even. People who talk of "the evolution of consciousness", and of its survival value, are obviously addressing third-person aspects of the problem. I don't mean to say that these aspects are uninteresting, but nevertheless these aspects are not what makes consciousness such a mysterious problem. It is a pity; for a while the word "consciousness" was a general indication that one was talking about the mysteries of the first-person. These days, this seems to be less often the case.

The cause of this confusion is of course the intimate relation that the first person has to the third person. We should never forget that the mind is caused by a brain, and that the brain is at the bottom line a physical system understandable from the third-person view. Although we do not know how, a first-person is emergent from a third-person-understandable substrate. A consequence of this is that much of interest from the first-person viewpoint *corresponds* directly to phenomena viewable from the third person. Even the thought which I am having now: "Wow, how could it be that a mere brain could *experience* this thought", is being supported by a pattern of neural activity in my brain.

Take "consciousness," for instance. Despite the fact that this word usually represents all that is mysterious about the first-person, it would be naive to expect that the phenomenon be completely separable from the third-person viewpoint. And indeed, there is much in the third-person viewpoint which gives us insight into consciousness. The third-person notion of a brain which is scanning itself, or observing (directly or indirectly) its own processing, for example, obviously has a lot to do with consciousness - it is an important part of the third-person substrate from which consciousness emerges.

But it would be a mistake to regard this view of consciousness as dissolving the first-person mysteries altogether (as is sometimes claimed). To make this distinction clear, I will always denote this view of consciousness as "third-person-consciousness", or in the interests of brevity, "3P-consciousness." (One should not confuse this third- -person view of consciousness with "consciousness of the third person," which is a different matter entirely.) When I use the word "consciousness" alone I will always mean "first-person-consciousness", which I will sometimes abbreviate "1P-consciousness."

This direct correspondence (some might even say isomorphism) between first-person phenomena and (a certain subset of) third-person phenomena seems to be what often leads to confusion when discussing first-person issues. Many commentators, particularly those in the third-person camp, give the illusion of reducing first-person mysteries by appropriating the usual first-person words to refer to the third-person phenomena to which they *correspond*. It would be a final irony if this was to happen to the word "first-person" itself. I hereby issue a plea that this word be off-limits to the third-personites. If they wish, they may argue that the first-person *does not exist*; but they may not pretend to 'explain' the first-person by describing only third-person phenomena.

It would be nice if every article on 'mind' and 'consciousness' came with a caveat at the beginning, alerting the reader whether it is to be the first-person or the third-person phenomena that will be discussed. It is not unusual to find a paper which seems to be addressing the great first-person mysteries, only for the reader to find halfway through that it is doing no such thing. Sometimes even authors themselves seem confused as to which questions they are addressing.

It seems that there are only about three expressions which these days are still reserved only for first-person phenomena. These are "qualia," "phenomenology," and "subjective experience." But even these words (particularly the last) may begin to be appropriated by reductionists; and besides, each of these words has a fairly limited domain of application. I will always use the term "first-person" as a general term covering the whole area of this metaphysical mystery.

Of course there is also the old standby "Mind-Body Problem." This phrase has probably outgrown its usefulness, with the change in usage of the word "Mind" and the change in emphasis from "body" to "brain." But I will still use it from time-to-time; no other phrase has such universal first-person connotations.

The Mystery of the First-Person

As I have said, it is difficult to talk about the first-person without descending into vagueness. But what can be done, if it is done carefully, is to point out the mysteries, and ask how a third-person, physical theory could ever deal with these. I do not intend to do this in this paper - I take it that this has already been done most ably, by Nagel and others, and that reductionists have never given an adequate response. To re-present these arguments would be to go over old ground.

Instead, I intend to find out what the third-person can tell us about the first- -person, and to present a first-person theory which coheres with the third-person view. Any hard-line reductionists reading this will of course not be convinced by my premise. They may read this paper perhaps for amusement, perhaps because these first-person ideas may give them some insight into their third-person ideas, and perhaps even to realize that not all believers in the first-person are as unreasonable as they had supposed.

I will, however, give a very brief account of some of the most difficult first person problems. Two of these stand out above the others.

(1) The problem of QUALIA.

Qualia are the qualitative aspects of subjective experience, particularly sensations such as colour, taste and pain. One can ask: how could a third-person theory begin to explain the sensation of seeing the colour red? Could a theory be given which enabled a being, as intelligent as ourselves but without the capacity for sight (or even a visual cortex), to truly understand the subjective experience of the colour red? Analogously, as Nagel asks, could we ever understand how it feels to BE a bat, or another creature with very different mental architecture.

It is extremely difficult to imagine how a physical explanation of brain architecture could solve these problems. If, in answer to the question "could a sightless being understand the sensation of red?", the reductionist answers "yes," then we have the right to ask "How, possibly?". I believe that no satisfactory answer to this has been given. If, on the other hand the reductionist answers "no," then the very least of her problems is a serious epistemological limitation.

(2) The problem of MENTAL CONTENT

This problem is harder to articulate than the previous one, but I believe that it is even more important. When we think, our thoughts have *content*. How is it possible that, in some absolute sense, patterns of neuronal activity in a small biological system support *meaning*? We can *stipulate* meaning to a physical system (as is done by Dennett's 'intentional stance', and perhaps by Fodor's 'language of thought'), but do these stipulations have any metaphysical reality? If it wasn't for the direct evidence of our introspection, the idea that a physical system could in an absolute sense carry meaning would be dismissed. And yet it happens.

There are other ways to articulate first-person problems. We could talk about the problem of "subjective experience," or of "consciousness," but we begin to tread on slippery terminological ground, which it is probably best to avoid if possible. In any case, all of the first-person mysteries are encapsulated in the term "Mind-Body Problem." People who think that this problem has been solved are surely in the minority.

{end of linear article so far...}

{ {fit this in somewhere?}}

Most writers on the subject of 'consciousness' recently have chosen to take a very third-person stance, talking about such things as the "Evolution" and "Architecture" of consciousness, while in fact talking about no such thing. They talk about fascinating third-person properties of thinking systems, which begin to explain why the systems produce these outlandish CLAIMS of consciousness and so on; and even explain how the ability of these systems to monitor themselves leads to the appearance of consciousness; but only very rarely is the bottom-line question of "how could ANY physical system support a first-person?" addressed.

One of the these of this article is that we should not be at all surprised at the success of these writers in explaining the "appearance" or evolution or "illusion" or achitecture of consciousness - claims of consciousness are something we should expect every intelligent creature to make, and these claims by themselves are not inherently mysterious facts. There is plenty of fertile ground for Reductionists to give theor numerous explanations of the "fiction of the self" and the "evolution of consciousness". But the first-person, subjective question is still not answered. Only rarely do reductionists take more than a paragraph or two to address the true concerns of the first-personites, and when they do it is usually to say "Well, we've explained why a physical system might believe in these fictions, and that's all we need to do." But if they don't answer the question of how a physical system night have semantic content or subjective experience, then this is wasted effort. [Aaaargh, rephrase this.] One reductionist who does a great job of applying third-person thinking to yield though-out first-person conclusions is Parfit, but he is definitely in

the methodological minority.

[Last two sections are horribly incoherent. Mass reorganization required.]}

Philosophers these days who argue for the mystery of the first-person view are often regarded as having a slightly anti-materialist bias. I try to show in this paper that this is not necessarily the case. I will present a materialist framework which, I believe, can *explain* the first person without *dismissing* it. One can be a materialist without being a hard-line reductionist.

...

I intend to present a framework for discussion of the first-person that still makes sense from a third-person view. The theory which I present is, as I see it, the only real alternative to the cult of the third-person. In fact I have a sneaking sympathy with the reductionists, and I believe they have a passing chance of being correct. Who knows, perhaps they are in fact right, and consciousness is just an incredibly powerful illusion. (Especially in view of the occasional really good argument such as that of Parfit). But this is far from demonstrated now. If a first-person theory is to have any chance of being coherent, I believe it must fit into the framework which I present.

A Taxonomy of Cognitive Jokes

[David J. Chalmers](#) (February 1989)

By a 'cognitive joke', I mean a joke whose humour seems to rely on higher-level, more abstract cognitive processing in the brain. These contrast with sexual, violent and other types of jokes which rely on our 'baser' and more instinctual reactions.

This is just a beginning categorization. I claim no 'objective correctness' for it. And of course the categories can be fluid, and the same joke can be a member of more than one category (and perhaps it will be funnier if it is). But thinking about the jokes which I can recall from the Humour Weekend, most seem to fall squarely into one or another category, indicating that perhaps this is a useful way of dividing jokes. It seems to me that the "causes of humour" in all 4 classes are different, coming from different parts of the brain.

The 4 basic classes...

(1) *Allusive*: Allusion to common situations or shared human foibles, producing a spark of recognition (from experience) when told.

(2) *Interpretational* (or Gestalt): 2 ways of interpreting exactly the same situation - or alternatively 2 situations that fit the same description. This is perhaps the largest category. It seems to arise from the power of the brain to interpret situations flexibly. This category is divided into many subclasses.

(3) *Paradoxical*: Typically, self-defeating statements or contradictory situations. Examples of this small class often have a flavour of mathematical logic.

(4) *Inferential*: The listener is forced to make an inference from information given in the joke (especially in the punch-line, which typically seems to "fall flat" for a moment), to reach a surprising conclusion. The pleasure seems to be mostly in the process of inference itself.

CATALOGUE

(1) Allusive

- Doonesbury, often (e.g. teaching strip, "everything about you")
- Russian "that reminds me of my cow" ;;this is also inferential
- Viola jokes
- Meaning of Liff
- Bohr "profound truth"
- Marriage jokes

- some ethnic jokes (good ones as opposed to generic ones)
- most political jokes

The humour here is usually in the recognition of a common experience from life, of a common type of person or event. (Often a peculiarity that is rarely commented on explicitly but recognized from everybody). Typically there is a 'spark' of recognition - "that is so true".

If the allusion is not recognized directly, then these can be taken as "information-conveying" jokes (Yan Zhao). In this case the jokes take on a mildly Inferential quality.

(2) Interpretational

There are various categories of interpretational jokes, perhaps the most common type of cognitive joke. All involve the notion that a situation can be matched to a description (or interpretation) in more than one way. This can happen in various ways: either the description or the situation can be given as the initial data, and the 'branching' can occur either in the step from situation to description or vice versa.

We classify these jokes by the order in which these events occur. Our classes are labelled SDS, DSS, SDD and SD.

SDS: These could also be called "Do this" jokes. A situation S1 is given, and typically somebody (in the joke) is required to perform the action in S1 themselves, in some sense. This is a kind of analogy problem, so a description of the situation must be formulated. Usually there is an obvious or expected interpretation D1, which would lead to situation S1' under analogy; but instead a surprising interpretation D2 is made, leading to an unexpected situation S2 (a 'false analogy').

(so SDS == situation -> (wrong) description -> unexpected situation).

- 1 - Garfield "do this"
- 2 - Copycat jokes
- 3 - lion-tamer joke

DSS: Here a linguistic description D of a situation is given. There is an obvious (to us) implementation of this description as situation S1. However, as it turns out, whether through stupidity or quirk of fate this description is actually implemented as a surprising situation S2. (DSS == description -> expected situation S1 AND surprising situation S2.)

- 4 - Poles changing lightbulbs
- 5 - Poles washing car-windows
- 6 - Alaska superconductor
- 7 - "wanted to look like Liz Taylor"
- 8 - Pearls and "The Meaning of Dreams"

- 9 - Reagan juror for North trial
- 10 - filing parrot's bill

These cases SDS and DSS are closely related. Both involve a situation S1 as expected implementation, and a surprising implementation S2 which occurs instead. The difference is whether the implementation is situation-based ("Do this", by analogy) or description-based.

There are some sub-categorizations which run across the border between these categories.

(1) The difference between the two situations S1 and S2 can be a reversal (figure-ground) relation (1, 3, 4, 5), a 'parallel slippage' relation (6, 7, 8), or neither (9, 10).

(2) Most of these cases are 'goal-directed' (except 8, 9). We can regard the description (or the "Do this") as a goal to be achieved. In all cases the goal is achieved in a surprising way. This can either be through stupidity (resulting in a harder than necessary solution) (4, 5, 10), through cleverness (resulting in an 'easy' solution) (1, 3, maybe 6, 7), or maybe through non-intentional quirk-of-fate, resulting in a 'different' and usually easier solution (6, 7). (Perhaps regarding 6 and 7 as goal-directed is stretching things somewhat.) (Incidentally, note the presence of Allusion in 6, 7, 9.)

We have to check how independent these variables are. (e.g. I believe SDS and parallel slippage may be incompatible, though this may be false.)

SDD: Here we start with a situation S, which has a correct or obvious interpretation D1, but which is instead perceived as the surprising interpretation D2 (usually mistakenly, else facetiously).

- "Is the doctor home?"
- Zero-legged frogs are deaf
- Pan-Am luggage labels.

These are closely related to SDS jokes. In both cases two descriptions are found to fit a situation. The difference is that in SDS jokes, these interpretations are implemented as analogical situations, whereas here the surprise is explicitly in the description.

SD: In this case we have a situation S which has NO obvious explanation until the punch-line. The punch-line supplies a weird but sense-making and 'correct' interpretation D.

- Superman at Empire State Building
- Sperm cell from mouth
- Bulldog, gorilla and shotgun.

These could possibly also be classed as Inferential jokes - see below. They are also quite closely related to SDD jokes. Also, some DSS jokes could be rephrased as SD jokes, by giving the surprising situation S2

first and waiting for the explanation. ("Did you hear that Reagan is a juror for the North trial")

DS: Newsflash from the front! Here is a new, similar category DS. A description is given that seems difficult to satisfy, and then a weird implementation S, which satisfies the description, is given in the punch-line.

- Many riddles, e.g. "How do you sink an Irish submarine?"

These have a similar feel to SD jokes, with the same "Inferential" flavour. They are also quite related to DSS jokes.

So we have (at least) 5 categories of Interpretational jokes. These are SDS, DSS, SDD, SD and DS jokes. With SDS, DSS and DS jokes, the humour is in the unexpected implementation S2 or S, which is typically revealed in the punch-line. With SDD and SD, the humour is in the unexpected interpretation D2 or D, again typically revealed in the punch-line.

As we have seen, there is an intricate set of inter-relationships between the categories. This could be explored in more detail.

The humour in all categories seems to be based on the exceptionally flexible mechanisms which human beings have for matching interpretations to situations.

(3) Inferential

Here typically, when the punch-line is delivered, a situation is set up which at first seems unexplained and perhaps incomplete. For a moment the joke seems to fall flat - "is that it?". Then, after a moment the listener is forced to make an inference from information present in the text to form an explanation, which is often surprising or weird. The pleasure seems to lie mostly in the actual process of inference, and in the discovery of the explanation.

- Shostakovich = Beethoven backwards
- deaf genie (12-inch pianist)
- soldier cuts off his other arm
- self-referential kangaroo joke
- Local-Hero-type jokes
- Russian cow joke

These jokes are quite closely related to Interpretational SD jokes, which often need mild inference to explain a situation. The difference is one of degree; in SD jokes the inference is usually very easy upon the punch-line, and the humour lies mostly in the surprising nature of the interpretation. If the inference in Inferential jokes were less difficult, many could perhaps still be mildly funny SD jokes; but for me at least, it is the inference which gives these jokes their special quality, and makes them typically much

funnier than an ordinary Interpretational joke.

(4) Paradoxical

This is a smaller class of jokes, which seem to be based around self-defeating statements and paradoxes.

- "Individuals...I'm not"
- Cut off right arm to be ambidextrous
- 1000000 times, don't exaggerate
- "Yesterday I could raise my arm THIS high"
- soldier who cuts off his second arm

Here the humour comes when we realize the contradiction. (Occasionally a process of inference is required, depending on how obvious the contradiction is, so this class is not totally disjoint from (3).) Perhaps the appreciation of these jokes comes from the "logical centers" of the brain, which have a relish for playing with paradox. Alternatively, the pleasure could be produced from the dissonance set up by the contradiction.

There are some other jokes with a somewhat logical framework.

- "Am I getting stronger or weaker?"
- God's stone - so heavy She can't lift it?

Analog vs. Digital Computation

[David J. Chalmers](#) (February 1989)

It is fairly well-known that certain hard computational problems (that is, 'difficult' problems for a digital processor to solve) can in fact be solved much more easily with an analog machine. This raises questions about the true nature of the distinction between analog and digital computation (if such a distinction exists). I try to analyze the source of the observed difference in terms of (1) expanding parallelism and (2) more generally, infinite-state Turing machines. The issue of discreteness vs continuity will also be touched upon, although it is not so important for analyzing these particular problems.

We find that to truly explain these problems we have to answer hard questions about the nature of the universe.

Analog solutions to hard computational problems. (e.g. NP-Complete ones).

Examples:

- 1) Spaghetti sorter - time n (as opposed to $n \log n$)
 - Cut n pieces of spaghetti in proportional lengths, bang them against the table.
- 2) String shortest-path finder - constant time (?) (vs ?)
 - Cut strings of the appropriate length between free-floating "vertices", then tug on the two appropriate vertices until there's a straight line.
- 3) Cog-turner to solve NP complete problem (?) in time n .
 - Described in Scientific American

These solutions all cheat by using PARALLEL computation, inherent in nature. But, you say: these problems can be shown to be just as hard on parallel machines as on serial machines (things just get speeded up by a constant factor).

But aha - there is something special going on here. Nature is not just providing us with an ordinary parallel computer. It is an EXPANDING parallel computer whose parallelism expands to meet our needs. In the spaghetti sorter, for instance, we essentially have n parallel processors, one for each stick of spaghetti. As n increases, so does the number of processors. This is unlike any typical digital computer in our experience.

Using natural parallelism, here is another, digital solution: Building n digital processors is only going to take time n (with a rather large constant factor). Once this was done then many of these hard problems would become easier. With a self-replicating machine, humans could even be taken out of the loop altogether!

Obviously both of these solutions are cheating mildly when compared with the original formulation. The original formulation only applied to nice, Turing-Machine-like algorithmic digital devices. The first solution uses "arbitrary" properties of the real world; and even the second "jumps out of the system" by physically fabricating new machines. But still... does this not mean that perhaps the original definition of 'algorithm' was a little restricted (Church and Turing turn in their graves)?

Anyway: the essence of the trick lies in EXPANDING PARALLELISM. We could give a formal definition of such machines and see what theorems apply to them.

Question: although we talk of (indefinitely) expanding parallelism, is it not the case that we start with one huge parallel machine to start with, namely the physical universe? We are not so much expanding our machine as making better and better use of its resources. If the universe is infinite, this is OK - improvement can go on indefinitely. But if the universe is huge but finite, then "expanding parallelism" is only an illusion. In the end we are going to exhaust all the spaghetti sticks in existence! This would be analogous to having a computer which was hard-wired to solve a problem extremely fast for $n < 100$, but which failed for all $n \geq 100$.

(General principle for all these analog 'solutions': Think hard about what happens for very large n .)

Basic question: is the universe an infinite-state-machine or a finite-state-machine? If the second, then it is WEAKER than a Turing machine, so these analog solutions are essentially weak theoretically. If the first, then is it true that it is STRONGER than a Turing machine? At the very least, it seems that it has different theorems of computational complexity.

The only way a Turing machine can accommodate 'parallelism' is through its states. But as a TM has by definition only a finite number of states, then indefinitely expanding parallelism is out of the question. (Of course, such parallelism can be 'simulated' serially by use of its tape, but this is not helpful for questions of computational complexity, although this is good enough for questions of computability.)

If we have an infinite number of states, it seems that indefinitely expanding parallelism can be accommodated. As we've seen, this leads to better results for questions of computational complexity. It is an interesting question whether such a machine was more powerful computationally (in terms of the problems it could solve, independent of resource-use).

Presumably we have to specify our definition a little better. First: do we allow uncountable number of states? This might behave even better, but presumably this is outside the spirit of computation, which has it that algorithms must be well-defined and in a sense 'constructible'. The well-known unconstructibility of uncountable sets (or at least of most of their members) would seem to mean that such a machine is inherently "impractical".

OK, given a countable number of states - we'll index them by the positive integers. Next question: what do we allow as our transition functions? Do they have to be computable functions, or do we allow any function $Z \times \dots \rightarrow Z$ to be 'hard-wired'. This question does not arise for finite-state machines, because here any function is computable. My intuition here is that again, to be computationally 'reasonable', such functions must be computable and maybe even primitive recursive.

Incidentally, where does the physical universe fall into these classes? Assuming it is (a) infinite and (b) discrete (this is arguable) then I believe that, although the number of 'possible' states is uncountable (from a calculation like 2-to-the-infinity), the number of 'reachable' states (from constructible initial situations) is countable. And given that the laws of physics are reasonable, the transition function is computable and indeed primitive recursive (though it may be non-deterministic).

My basic idea of an infinite-state-machine is like this. As we already have an infinite number of states, we have no need for a tape - we can simply incorporate this into the states. One obvious way of specifying a state is by an infinite string of 0's and 1's. (Note that this gives an uncountable number of states, however). We simply require that every string at time t determines another string at time $t+1$.

(In imagining this, we imagine ourselves 'outside' the system and determining its initial state, or at least a relevant portion of it. Of course in the case of the physical universe we are in fact INSIDE the system; but given the 'illusion' of free will we can treat it as if we were manipulating things from the outside).

The idea of having "reasonable" laws of physics corresponds to the idea of having a computable transition function. So I require that the digit at position P in the string be a primitive recursive function of the previous string. (To make this totally meaningful, we may have to make it determined LOCALLY; a digit P can only depend on its N nearest neighbours.)

This definition corresponds fairly closely to an infinite discrete universe. Another possibility would be the 'potentially infinite universe'. Here we require that all initial configurations consists of 0's outside some finite range. The 'local' nature of the laws dictate that information would always be contained within a finite range, but this finite range could be always expanding. Perhaps, given the Big Bang theory, this corresponds closest of all to our universe. The computational power here is not obvious. It seems that indefinitely expanding parallelism is still accomodatable in such a universe, but my intuition is that it should be computationally weaker.

Incidentally, the potentially infinite universe is nice also because it only requires a countable number of states. I believe that this can be shown to be equivalent to the infinite-state machine where states are indexed by integers (a few paragraphs above.)

If in fact it turns out that the universe is non-discrete, it would take a strong argument to convince me that this led to inherently different results to a discrete universe. Everything we have seen indicates that all interesting properties of non-discreteness can be simulated using low-level discreteness, especially using non-determinism. But I may be wrong here. A finite but non-discrete universe might be worth analysing. As we've seen, a finite discrete universe is equivalent only to a (rather big) finite state automaton, so that despite our initial illusion of powerfulness, it is in fact weaker than a Turing Machine. If the universe was finite and non-discrete (as may well in fact be the case), then on the face of it, this is a powerful infinite-state machine, more powerful than a Turing machine. The question is whether the continuity on smaller and smaller levels can even theoretically be utilized by us computationally, or whether (for instance) randomness, 'chaos' and noise would make this system no more powerful than the discrete case.

Summing up, we have a few cases.

Infinite discrete universe: this is an infinite-state machine, with an uncountable number of states. It seems that it can accomodate indefinitely expanding parallelism. Results in computational complexity here are less limitative than those about ordinary Turing machines. It is not initially obvious whether computability power is greater or equivalent.

Potentially infinite discrete universe: this is an infinite-state machine with a countable number of states. This probably can accomodate indefinitely expanding parallelism, although the computational power may be a little weaker than above.

Finite discrete universe: this is equivalent to a finite-state automaton, so it is weaker than a Turing machine despite the initial impression of computational power.

Finite continuous universe: harder to analyze. Perhaps equivalent to the finite discrete case, perhaps more powerful.

Infinite continuous universe: could this be more powerful than the infinite discrete case? This is hard to say. The discrete case was extremely powerful already. Perhaps once we have infinitude already, then continuity can give us nothing extra.

The question of whether continuity is inherently more powerful than discreteness is a fascinating one which needs to be analyzed. Here, however I have mainly been concerned with issues of finite vs. infinite, serial vs parallel - these are the issues which underlie the 'powerful' analog machines mentioned at the start. Discreteness vs continuity may be another important aspect of the analog vs digital issue, but there is no hard evidence for

How Cartesian Dualism Might Have Been True

[David J. Chalmers](#) (January, 1990)

We could have been characters in a huge computer simulation. It is a familiar idea that the whole world might be simulated on a computer, and things would seem exactly the same to us (and indeed, who is to say that we are not).

I imagine, though, a different sort of simulation, of the kind common in the fields of artificial intelligence and artificial life, where we have (1) a simulated environment and (2) simulated beings which are "moving" through this environment, according to a program that models these beings' thought processes and their decisions. Imagine a very complex project like this (like the vivarium, say), perhaps with genetic algorithms which get more and more complex and sophisticated, until eventually very sophisticated, rational beings evolve.

When they speculate about the world, they will find that the environment possess certain regularities, and this will lead them to laws of "physics" about their external world. This will lead them to speculate about whether they too, at the bottom line, are subject to the same laws. This might seem plausible...but of course it will not be the case! Their "mental" life obeys a completely different set of laws, and further these laws are off limits for direct observation. Their mental life takes place not within their world at all, but within in a computer in a completely different universe! When it comes to observing the "laws" of their behaviour, they will reach some dead end in looking for causal mechanisms. Unlike our world, such mechanisms are simply not "locally supported" by simple physical laws. I'm not quite sure what would happen next.

If they tried to "look inside their heads" (assuming they have at least vaguely coherent senses)... They'd just find an empty box. They'd ask "how can I do all this complex processing". The answer would *have* to be, well, I'm just kind of non-material mind. Of course, there would be a breakdown in the usual kind of physical causation around the "heads" of such a being, unlike our world.

Moral: Cartesian Dualism isn't quite so outlandish and conceptually problematic as tends to be supposed.

-
- For a non-Cartesian form of dualism see my [Facing Up to the Problem of Consciousness](#).
 - Return to my [Miscellaneous writings](#) page.
 - Return to [David Chalmers' home page](#).

Deep Systematicity and Connectionist Representation

[David J. Chalmers](#)

Commentary on Andy Clark, "Theoretical Spaces"

1. I think that by emphasizing theoretical spaces of representations, Andy has put his finger on an issue that is key to connectionism's success, and whose investigation will be a key determinant of the field's further progress. I also think that if we look at representational spaces in the right way, we can see that they are deeply related to classical phenomenon of systematicity in representation. I want to argue that the key to understanding representational spaces, and in particular their ability to capture the deep organization underlying various problems, lies in the idea of what I will call *deep systematicity*. This is closely related to classical systematicity, but I think that it may potentially be much more far-reaching.

So I'll spend a few minutes developing this point, starting with a look at classical systematicity.

If we look at the cluster of phenomena that Fodor and Pylyshyn emphasize in their discussion of the Classical approach - compositionality, systematicity of cognitive capacities, productivity - we can see that all these derive from one key property of Classical representation: that the semantic structure of what is represented is systematically reflected in the formal structure of the corresponding representation. This is the general property that I'll call Classical systematicity.

e.g. look at [example and explanation]. we can see how this systematicity of representation immediately buys the phenomena we were talking about, and most importantly of all, it buys a *systematicity of functional role*. Representations that have a relevantly similar semantic structure, will, by virtue of their systematic form, play a relevantly similar functional role. So the propositions "John loves Bill" and "John loves Mary" will play similar functional roles within the system, with any differences being a systematic reflection of the difference in one constituent; and these two propositions will play an entirely different functional role to that played by a proposition such as "The quarterback fainted last Saturday."

This is guaranteed precisely by the complex internal structure in each representation. If the representations were structurally atomic, any such systematicity of functional role would have to be arbitrarily imposed by external rules.

It's difficult to overemphasize the power of this idea, and it's been at the root of much of the last century's work in formal logic and AI.

BUT...

are there semantic primitives?

classical atomic representations lack any internal structure

so there is no reflection of semantic similarity among such representations

inferences performed on such atomic representations are therefore quite unsystematic.

consequences include the well-known *brittleness* of symbolic AI models.

Reason: internal structure of a classical representation reflects at best the *compositional* properties of what is represented. but there are a lot more semantic properties out there.

2. This is where connectionism comes in, and connectionist representational spaces in particular. I think what ought to be regarded as the key philosophical commitment underlying connectionism is the following:

No structurally atomic representation.

Instead, in connectionist systems, *every* representation has complex internal form. This is made possible by the universal use of *distributed representation*, where the representation of any concept is spread over a number of separate computational units, none of which is a representation in its own right.

Once we allow every representation to have complex internal structure, then we have opened the door to what I will call *deep systematicity*, that is, the reflection of semantic properties of what is represented in the formal structure of the corresponding representation, *all the way down*.

In connectionist representational spaces, that's just what we find. After a period of learning, representations fall into a multi-dimensional vector space, such as the one depicted here (in a diagram lifted from PC). What we find is that the structure of this space is in no way arbitrary. Instead, objects that are semantically similar, in a way that is relevant to a given problem, have corresponding representations that are formally similar - lying close together in this vector space, for instance. In a connectionist network that classifies animals, for instance, the representations of "cats" and "dogs" will be much closer together than those of "cat" and "elephant". This is a direct consequence of the fact that relevant semantic properties of the "cat" concept are reflected directly in the structure of the vector that represents "cat".

Of course, this space here is just one of Andy's theoretical spaces, that, assuming it is successful, has developed in a way that captures the deep organizational structure of a given problem space. So theoretical spaces and deep systematicity are two sides of the same coin.

Now what does deep systematicity buy you:

(1) Firstly and most importantly, it buys the key property of *automatic generalization* that Andy has been stressing. If a representational space has been built up that handles cats and mice and elephants, and all of a sudden a hamster comes along, then deep systematicity causes the hamster to be slotted into a vectorial representation that is likely to be closer to the representation of "mouse" than that of "cat", and certainly than that of "elephant", though also different in relevant ways. Any actions to be taken will depend directly on this representational structure, so it is likely that if we've got the representation right, we should get the action about right too - and that's what we find in practice.

Other properties that come along with deep systematicity includes the ability of *context* to subtly influence the form of a representation, and therefore the consequent actions - something that is more difficult to achieve when objects are represented by simple primitives; and also the ability to interpolate between two different representations when appropriate.

Now, of course the existence of a deeply systematic representational space doesn't do *all* the work for you - you still need processes that will operate on the representations, allowing them to play appropriate functional roles. But the point is that (just as in the case of Classical systematicity), the systematicity in representational form provides a strong assurance of systematicity of functional role - i.e., a functional role that will cohere with the semantic properties of what is represented. In effect, deep systematicity tries to, as far as possible, take the responsibility for reflection of semantic regularities out of the rules, and into the representations.

Words on classical systematicity (tensor products, RAAM).

Now, to get back to the problems that Andy was worrying about, it seems to me that the central problem is just this: How can deep systematicity be *achieved*? Connectionism per se, through the introduction of distributed representation, *allows* the possibility of deep systematicity. But that potential needs to be backed up by ways of getting there - by actual mechanisms that develop deeply systematic representational spaces. So far, connectionism has developed just a few algorithms, that demonstrate the potential, but they're far from exhausting the possibilities.

In connectionism's most frequent incarnation to date, deep systematicity of representation is achieved through the use of the backpropagation algorithm. Now backpropagation is terrifically powerful, but Andy has criticized it, rightly I think, for being too tightly tied to first-order properties of the input and output spaces. If the relevant organizational properties of the relevant representational space can be straightforwardly extracted from the input and output spaces, then it will do fine, otherwise it tends to have trouble. The problem seems to me to lie in the fundamentally associationist nature of the algorithm. What is being learned is an input-output association, and the hidden representation is constrained to be a simple way station on the path from input to output. There is little room for the complex processing that might exploit more abstract properties of the input and output spaces when necessary.

Andy's response is to lean on the possibility of exploiting the sequential order in what is learnt. I think

this is an intriguing possibility, but that he may be overestimating it. For even incremental learning shares with traditional backpropagation a strong associationism - indeed backpropagation is the algorithm that Andy supposes is used throughout. So it seems to me that the associated problems will not so easily be avoided. It may well be that the right kind of incremental learning will provide a slight improvement over vanilla backpropagation, to speak somewhat figuratively an improvement of 20 or 30 percent; but it seems to me that it's not the kind of thing that will lead to a vast qualitative difference, on the order of 200 percent. Incremental learning or not, backpropagation remains strongly driven by the first-order properties of input and output spaces. The empirical evidence is not yet in, but I've heard of as many failures with incremental learning as successes, so I think there are grounds to be skeptical.

Andy's proposal shares with traditional connectionism the tendency to lean as strongly as possible on *learning*. There seems to be a deep suspicion in these circles of positing too many innate mechanisms. To some connectionists, to posit that a problem is solved by the use of innate mechanisms is something of a cheat, if it could be solved by learning instead. You might say that on this view, an organism has to *earn* its abilities.

I don't entirely agree with this attitude, but neither do I agree with the attitude that the only real explanation of systematicities must be due to innate architectural constraints, an attitude that is suggested by the work of Fodor and Pylyshyn. This view goes to the other extreme, arguing that innate mechanisms carry the real burden, and that so-called "learning" is just a matter of the triggering of the right innate mechanisms.

Instead I think there is a middle ground that allows the admission of strong innate mechanisms while retaining the connectionist emphasis on the importance of learning. This view, which I've objectively labeled "enlightened connectionism", sees our cognitive capacities as the result of the continuous adaptive processes of evolution and learning.

Of course there's nothing at all radical about this - it's just a way of looking at the development of cognitive capacity that causes all of the nativist vs. empiricist problems to seem relatively unimportant. It involves the recognition that evolution and learning are processes of adaptation that are relevantly similar in kind: both process whereby organisms can climb up adaptive landscapes, improving their ability to deal with an environment. Evolution takes you so far, and learning gets you the rest of the way. That point that is so important to nativists and empiricists alike - that is, the question of what is built-in at birth - is reduced to the status of a relatively unimportant border where these two processes of adaptation meet. Furthermore, this line between evolution and learning is actually quite arbitrary. There are studies showing that what was once achieved by learning, some time in a species' past, can later be achieved by evolution. It's just a matter of which is more efficient.

With this view in hand, we can construct a view of cognitive development that is quite compatible with the spirit of connectionism, but which eschews any simple form of empiricism. For a start, the only tabula rasa are in the priordial soup. Everything since then has been the result of an adaptive process, and so is very much a product of its evolutionary environment, with mechanisms developed that are adapted

to that environment. On the other hand, we can't throw around innate mechanisms willy-nilly - on this methodology, there at least has to be some reasonable evolutionary explanation of how an innate mechanisms might have come about.

As a very rough rule of thumb, we can say that in the development of cognitive capacities, evolution may be likely to build mechanisms that exploit regularities that are robust across a number of different environments, to deal with problems that any member of a species is likely to have to handle; whereas learning is likely to handle regularities that are specific to the environment that an individual organism finds itself in. If we apply this to the case of the classical systematicity of cognitive capacities, as Andy considered at the end of his talk, we can see that it may be more plausible that the mechanisms responsible for this systematicity might be a product of evolution, precisely because it seems to reflect a similar robust systematicity in most environments. One doesn't just find trees to the left of lakes; one also finds lakes to the left of trees, and so on for a very large number of relations and objects, though by no means all of them. So it maybe more reasonable to suppose that classical systematicity would be sufficiently adaptive that it might be built in by an evolutionary process. Of course, as we've seen, the line is somewhat arbitrary - it is also quite possible that some of the job might be left to learning, as long as it was guaranteed that learning could do the job relatively quickly, and reliably across different environments. The latter suggestion might also help explain the gaps in systematicity that are purported to exist in at least some animals. But it does not seem to me that systematicity should entirely be a product of a difficult process of knowledge acquisition, as Clark suggests - that would seem to be a maladaptive way to deal with such a vital capacity.

All of this would come to very little if it were not for the fact that there exists a perfect computational tool to go along with this variety of connectionism, and that is the genetic algorithm.

[explanation goes here]

Now this can be combined very directly with the use of connectionist networks, e.g. to specify the modularity of a network, or the topology, as an initial state. Of course we only specify the state of an organism at the beginning of its life; we then allow learning to take over. Those organisms are most fit that are doing the best *after* a period of learning.

What we find is that individual organisms are selected that have a complex initial structure that allows them best to learn problems with given domains.

How does this relate to deep systematicity? Well, the explanatory burden need not fall entirely onto learning. Instead, we can find - and experiments have been done along each of these lines, that through an evolutionary process, we can develop innately:

- modular structures within network design
- biases upon the learning algorithms that networks use
- development of appropriate learning algorithms for appropriate domains

- initial weights that are well-suited to particular domains
- etc.

There have been various studies showing how connectionist learning and generalization can be improved by each of these techniques. These might seem to be "innate" mechanisms that a connectionist might like to rule out - Andy talks of building in "task-specific information" -- but in fact we are nowhere buying something for nothing. We need only suppose that a given task, or more likely a given class of tasks, has been around in an organisms evolutionary environment, and it is quite natural that mechanisms might develop that are specifically to handle certain classes of tasks, or even more likely, to enable the right organizing structure of given task domains to be picked up quickly and easily.

In this way we can step back from the empiricist/nativist debate, and admit a strong role for evolution in the development of such cognitive capacities as deep systematicity while still admitting the centrality of learning in cognitive explanation.

Determining the Moment of Consciousness?

[David J. Chalmers](#)

Commentary on Valerie Hardcastle, "Determining the Moment of Consciousness".

It's very interesting to see neurophysiological evidence brought to bear on the puzzling question of conscious experience. Many have observed that information-processing models of cognition seem to leave consciousness untouched; it is natural to hope that turning to neurophysiology might lead us to the Holy Grail. Still, I think there are reasons to be skeptical. There are good reasons to suppose that neurophysiological investigation contributes to cognitive explanation at best in virtue of constraining the information-processing structure of cognition. Of course this is a very large and significant role for it to play, but it may be over-optimistic to suppose that it can play some further explanatory role, taking us where information-processing theories cannot. If so, then neurophysiological accounts will be no more and no less successful at dealing with consciousness than information-processing accounts are.

I will elaborate these considerations later. For now, I will focus on the problem at hand: the difficult questions about the timing of conscious experience raised by Dennett and Kinsbourne. It seems to me that these considerations apply equally to this problem. If an information-processing account of cognition is insufficient to determine a fact of the matter about certain conscious experiences, then a neurophysiological account will not help us. If a neurophysiological account does the job, an information-processing account would have worked equally. In what follows, I'll examine the interesting data about evoked-response potentials that Hardcastle has presented, and see whether it helps in determining the various facts that Dennett and Kinsbourne claim are indeterminate.

First, it is useful to get clear on what is going on. There are two experimental scenarios that are relevant. In the first, a prime - say "CAT" - is presented for a lengthy period of 500 milliseconds, so that the subject is aware of having seen it, and shortly thereafter a target - say "DOG" is presented. The semantic relation between prime and target has a significant effect on reaction time, but more relevant for our purposes is an evoked-response potential that is consistently found to occur. This is the N400 evoked-response potential, starting about 300 msec after the presentation of the target. In the second scenario, a mask is presented followed by an extremely brief presentation of the prime, for only 33 msec, and then the target is presented. In these conditions, the subject is unaware of having seen the prime, but reaction time is still facilitated. Furthermore there is a separate anterior negative effect, the so- called "Ne" effect, around 200 msec after target presentation. In these experiments, the N400 potential is associated with access to explicit memory - that is, to information that a subject can explicitly recall - whereas the Ne is associated with implicit memory - information that a subject cannot explicitly recall, but which affects behavior in various subtle ways. These associations have appeared in other experimental conditions, leading many to suppose that N400 is associated with a separate system for explicit memory and that Ne is associated with a system for implicit memory system, perhaps specialized for lexical information.

Let us take this association of ERP's with memory systems as granted. What light does this shed on the

difficult questions about conscious experience?

First, I take it that the importance of masked priming experiments to Dennett and Kinsbourne lies in the following question: was the masked prime consciously experienced? On the face of it, there are two possibilities: the so-called "Orwellian" hypothesis, that it was briefly experienced but then forgotten, and the "Stalinesque" hypothesis that it never made it to consciousness at all. Dennett and Kinsbourne claim that there is no way to decide between these hypotheses, and indeed that there is no fact of the matter between them. Do the results under consideration help?

It seems that they do not. The results are compatible with both the Orwellian or Stalinesque hypotheses. What we are told is that information about the masked prime is stored in an implicit memory system; one to which the subject has no conscious access. But this tells us, at best, that the memory is unconscious; it tells us nothing about whether the perception of the prime, way back when, was conscious. Perhaps it was consciously experienced but never made it into explicit memory, only into implicit memory; this is the Orwellian hypothesis. Perhaps it was never consciously experienced, but was perceived subliminally and found its way into the unconscious implicit memory system; this is the Stalinesque hypothesis. Either way, upon later presentation of the target we would expect activation of the implicit memory system but not the explicit memory system, leading to the observed evoked-response potentials. No results about the triggering of memories can possibly distinguish the Orwellian and Stalinesque hypotheses, which by their nature differ only on events around the time of perception. It follows that the neurophysiological timing of memory retrieval is irrelevant to this question, and that the experimental results do not distinguish between the Orwellian and Stalinesque hypotheses. Dennett and Kinsbourne's indeterminacy remains undetermined.

I take it that Hardcastle is not claiming that her results settle this indeterminacy; that is, that they settle no questions about the experience of the prime. Rather, the claim is that they settle some questions about the experience of the target.. It is somewhat unclear just what the question to be settled is; certainly, there are no obvious Orwellian/Stalinesque mysteries hanging in the air for this case, in which the stimulus is perceived and remembered in a straightforward veridical manner throughout. The key claims that Hardcastle makes for the data are (1) that they suggest that implicit memory systems are activated before explicit memory systems, with a possible processing path from the first to the second, and (2) that they may allow us to determine the moment of consciousness in a fine-grained manner.

Looking at (1) first: Of course it is dangerous to infer from these localized measurements of potentials that implicit memory is activated before explicit memory, especially if both systems involve processes that are distributed in time and space; perhaps the potentials detect the tail end of the explicit memory retrieval, but the very beginning of implicit memory activation. But in any case, I don't think that isolating a processing flow from implicit to explicit memory systems would contradict any of Dennett and Kinsbourne's claims. As they say at the end of the BBS article:

There is nothing theoretically amiss with the goal of acquiring precise timing information on the mental or informational transactions in the brain. It is indeed crucial to developing a good theory of the brain's control functions to learn exactly where and when various informational streams converge, when

"inferences" and "matches" and "bindings" occur. But these temporal and spatial details do not directly tell us about the contents of consciousness. The temporal sequence in consciousness is, within the limits of whatever temporal control window bounds our investigation, purely a matter of the content represented, not the timing of the representing.

The contents of our experience of the target, I take it, are not here a matter of dispute. This leaves as the remaining live issue the moment of our experience of the target, as the title of Hardcastle's paper suggests. I take it that Hardcastle's goal is to establish the time at which some content becomes conscious. It is not clear to me that "the moment of consciousness" is even meaningful, but even if it is, the present data fall well short of establishing it. At best, we have timed certain specific physical events associated with a conscious experience. Presumably there are many such events, distributed in space and time, so it is unclear why this one - a fairly arbitrary potential that happens to be accessible to our measurements - should qualify as the moment.

This is made even clearer when we think about just what these potentials represent. We have been led to suppose that they correspond to the activation of certain explicit and implicit memory systems, carrying information about past events; but what does the time of activation of these systems have to do with the time of consciousness of the present perception? Very little, it would seem. Perhaps activation of these precedes "the moment of consciousness"; perhaps it comes afterwards. We might as well measure the moment of retinal stimulation, or the moment of verbal report. Either way, the "moment of consciousness", even if it is determinate, remains undetermined.

It seems that the neurophysiological data under consideration do not help us in settling questions about the moment and content of consciousness. There are good reasons why we should have been skeptical all along. In general, when one sees neurophysiological properties invoked in cognitive explanation, they are generally invoked in virtue of their contribution to certain information-processing properties. It follows that the neurophysiological properties give us little in the way of extra explanatory grip. One can see this in various neurophysiological "theories" of consciousness that have been put forward. For instance, Crick and Koch put forward certain 40 Hz oscillations in the cerebral cortex as the secret to consciousness; but upon close examination, these oscillations are put forward only in virtue of the role they may play in the binding of certain represented contents.

If we don't know why binding gives rise to consciousness, telling a story about the oscillations does not help; the other way around, if we have a good explanation of why binding gives rise to consciousness, the neurophysiological details are just the icing on the cake.

In Hardcastle's discussion, certain evoked-response potentials are invoked as explanatorily relevant. Here, it turns out that these are invoked precisely in virtue of their association with implicit and explicit memory systems. A detailed information-processing account of implicit and explicit memory would have worked just as well or as badly in settling the problems at hand; the neural properties give us nothing extra. Presumably Dennett and Kinsbourne would claim that timing information about explicit and implicit memory systems doesn't solve the problems about the timing of consciousness. Information about the ERP's is relevant here at best for the information it gives us about the timing of such memory

systems; searching for extra information in the ERP timing is therefore futile.

I'll conclude by considering the question of the "disappearance" of consciousness from information-processing accounts, and asking whether neurophysiology helps us at all. It seems to me that this "disappearance" comes to one of two things. First, the models provide no explanation of consciousness. Second, they provide no explanatory role for consciousness. I'm sympathetic with both these concerns, but I don't think that neurophysiology helps.

On the question of explanation, the problem is: whenever we are given an information-processing account of cognition, there always seems to be a further unanswered question: why should that sort of processing give rise to conscious experience? Nothing internal to the information-processing account answers this question, and consciousness therefore goes unexplained by such an account. This "explanatory gap" is not closed any further by a neurophysiological account. Given some neural correlate of consciousness, such as a N400 potential or a 40 Hz oscillation, we can always ask "why should that be accompanied by consciousness"? The only answer to this question that can usually be given appeals to the information-processing role that such a neural process plays - the role it plays in binding or in memory, for instance. But this gets us no further than before: the question of why that sort of information-processing should give rise to consciousness goes unanswered. It seems that on both the information-processing and neurophysiological accounts, we have at best an association between underlying properties and conscious experience, not an explanation.

The same goes for the question of the explanatory role of consciousness. Strange but true, it seems that information-processing accounts are providing good explanations of behavior even without their coming to grips with the mysteries of consciousness. To many, it seems that such accounts are logically compatible with the absence of consciousness. If so, then under certain natural assumptions, consciousness has no explanatory role to play in such accounts. This matter is controversial, but let us take the "disappearance" for granted, and ask again: does neurophysiology help? Again, it seems not. At best, neurophysiology will yield an association between neural processes and conscious experience. Perhaps we can take the questionable step of "identifying" the process and the experience, thus giving the experience an explanatory role by fiat; but again, this gives us nothing that an information-processing account could not. Certainly there are all sorts of associations between information-processing properties and conscious experience; take the association between consciousness and reportability, for instance. If we wanted to, we could identify consciousness with some associated information-processing property, thus giving it an explanatory role. The move is equally sound or equally dubious as it is in the neurophysiological case. I think there are good reasons why the move is unsatisfactory, but for present purposes we need only note that whether it is satisfactory or not, neurophysiological and information-processing accounts are on a par with each other.

Whichever way we cut the cake, consciousness seems to "disappear" from neurophysiological accounts of cognition no more and no less than it disappears from information-processing accounts. I conclude that the path to the Holy Grail lies elsewhere.

What is it like to be a Thermostat?

[David J. Chalmers](#)

Commentary on Dan Lloyd, "What is it Like to Be a Net?"

The project that Dan Lloyd has undertaken is admirable and audacious. He has tried to boil down the substrate of information-processing that underlies conscious experience to some very simple elements, in order to gain a better understanding of the phenomenon. Some people will suspect that by considering a model as simple as a connectionist network, Dan has thrown away everything that is interesting about consciousness. Perhaps there is something to that complaint, but I will take a different tack. It seems to me that if we apply his own reasoning, we can see that Dan has not taken things far *enough*. When we have boiled things down to a system as simple as a connectionist network, it seems faint-hearted to stop there, and perhaps a little arbitrary as well. So I will take things further, and ask what seems to be the really interesting question in the vicinity: what is it like to be a thermostat?

A quick glance at diagrams of the models shows that there are a lot of similarities between connectionist networks and thermostats [diagrams]. Both take an input, perform a quick and easy nonlinear transformation on it, and produce an output. Of course, there are a few extra units and connections in the connectionist network, but one wonders how relevant this whiff of complexity will ultimately be to the arguments about consciousness. Someone once said that there are no reasonable numbers between one and infinity. Once a model with five units, say, is to be regarded as a model of consciousness, surely a model with one unit will also yield some insight.

Indeed, if we apply Lloyd's own reasoning, the thermostat does very well as a model of consciousness. For a start: like NETTALK, it can be provided with an interpretation that also construes it as a model of the world. Specifically, it can be naturally construed as representing the world's temperature. Furthermore, it satisfies Lloyd's "coincidence condition", requiring a structural similarity between the world for the system and the world for us. Just as NETtalk captures salient distinctions in the world of English phonetics, the thermostat captures salient distinctions in the world of temperature. Indeed, it captures the single *most* salient such distinction, the distinction between hot and cold.

Now, to be sure, the thermostat does not use distributed representation, which Lloyd makes much of, and consequently it does not support the relevant superposition of information. To get around this disanalogy, we might relax our standards of simplicity just a touch and move to the "superthermostat", also known as the two-unit connectionist network [diagram]. It seems to me that we could achieve the kind of distribution and superposition that Lloyd is interested in with this network, but I am inclined to suppose that we need not clutter our model with such needless complexity. After all, does it not seem that this rich superposition of information is an *inessential* element of consciousness? To be sure, in the glory of human consciousness we find this richness, as Lloyd illustrates with the manifold aspects of a Tudor cottage, but the aim of the current exercise is to simplify and abstract.

Surely, somewhere on the continuum between systems with rich and complex conscious experience and systems with no experience at all, there are systems with *simple* conscious experience. A model with superposition of information seems to be more than we need - why, after all, should not the simplest cases involve information experienced discretely? We might imagine a traumatized creature that is blind to every other distinction to which humans are normally sensitive, but which can still experience hot and cold. Despite the lack of superposition, this experience would still qualify as a phenomenology. So, just as a connectionist network qualifies as a model of superimposed phenomenology, the thermostat seems to qualify as a model of this basic kind of phenomenology. Indeed, if we accept Lloyd's approach, it seems to follow that in this model we have stripped down the substrate of phenomenology to its bare essentials. (They say that Euclid alone has looked on beauty bare, but surely this comes close.)

Having come this far this quickly, perhaps it is time to sit back and get some perspective. A thermostat, or indeed a simple connectionist network, as a model of conscious experience? This is indeed very surprising. Either there is a deep insight somewhere within Lloyd's reasoning, or something has gone terribly wrong. I am inclined to think that there is an element of truth in both these diagnoses of Lloyd's counterintuitive claim.

First, the insight. What Lloyd's approach brings out is that when we try to isolate the kind of *processing* that is required for conscious experience, the requirements are remarkably hard to pin down, and a careful analysis does not throw up processing criteria that are more than minimal. What are some reasonable-seeming *functional* criteria for conscious experience? One traditional criterion is *reportability*, but this is far too strong to be an across-the-board requirement. It seems reasonable to suppose that dogs and cats have conscious experience, even in the absence of an ability to report. A weaker criterion is *introspectability*: perhaps for a content to be experienced a system needs to be thinking about the content, or at least able to think about the content. I am sympathetic, though, with Lloyd's remarks that this sort of thing seems to be more symptomatic of reflective consciousness than of primary consciousness. On the face of it, it seems plausible that we can experience the fringes of our visual field without thinking about those experiences; and it is not implausible that a dog, say, might have visual experiences but entirely lack the conceptual capacity to monitor those experiences at a higher-level. I don't think the higher-order thought view of consciousness leads to an infinite regress, as Lloyd suggests - on such a view, a first-order thought will be conscious if it is accompanied by a higher-order thought, but the higher-order thought need not itself be conscious, so the regress is terminated - but it nevertheless seems more appropriate as an account of reflective rather than primary consciousness. Of course the issue is complex, but strong intuitions suggest that a system could be experiencing while only thinking about the world, not about its own mental states.

Sometimes it is suggested that conscious experience is a consequence of sufficient *complexity*, but this answer simply slides over the problem. Complexity is often relevant to the existence of some high-level phenomenon, but what is relevant is never complexity *tout court*, but the role that this complexity plays in a system. Life requires complexity, because complexity is required for adaptation and reproduction. If complexity is required for consciousness, it will be in virtue of some further functional property that this complexity enables, and we are seeing that this functional property is hard to pin down.

Indeed, on reflection it is hard to see why the intuitions that lead us to ascribe conscious experience to dogs and cats should prevent us from ascribing it to mice, or to flies, or to simpler systems. Such systems may lack such frills as language, a rich conceptual system, the ability to introspect, and perhaps even a concept of self, but why should *experience* require any of these? When we look at the kind of processes that give rise to experience in humans, such as the processes underlying color vision, what seems most essential is that the processes make certain discriminations, and make the relevant information available to the overall system in the control of behavior. This is a very basic sort of information processing, and is something that even a very simple system might share. It is this sort of reasoning that makes it seem just possible, after all, that there might be something it is like to be a net, or even a thermostat.

I confess that I was a little disappointed, though, after Dan's bold title, to see his nerve failing at the crucial juncture. Almost as an aside, he concedes that in fact NETTALK lacks experience, and that the answer to his title question is therefore "nothing". NETTALK, then, is not an instantiation of conscious experience; it is only a *model* of it.

This claim is weaker on the surface, but in fact I find it even harder to believe that NETTALK is a *model* of experience than that NETTALK *has* experience, counterintuitive though that second claim may be. For a model carries a particular burden: it must *explain*. And this leads us to what is perhaps the central worry about Lloyd's approach. On the face of it, this approach is put forward as a way of dealing with Nagel's worries about consciousness, where the central mystery is: why is there something it is like to be us at all? There is a huge *prima facie* mystery about how any sort of physical system could possess conscious experience. Lloyd holds out the promise that connectionist models might shed light on this question, but at the end of the day the models seem to leave the key *explanatory* question unanswered. Even if we were to go out on a limb and suppose that these simple systems *are* conscious, the question of explanation would still remain untouched.

What is it that these models might *explain*? On the face of it, they hold out the promise of explaining our abilities to make certain distinctions, and to exploit those distinctions in the control of behavior. We might ultimately see how the formation of a sophisticated world-model through information-processing enables a repertoire of actions that reflect the sophistication of that model. But where, in these models, is an explanation of *experience*? The problem in explaining experience is the apparent gulf between the brain and the quality of experience itself. A *model* might be expected to make this link more intelligible, but these models leave it as wide as ever. Why should *this* sort of processing be responsible for experience? There is no answer to be found by examining the models, or indeed within Lloyd's discussion. These are only models of *processes*, and they leave the gulf between process and experience as wide as ever.

Perhaps the best way to regard these models is as follows. We take the existence of experience for granted, and note that at least in familiar cases, there is a remarkable coherence between the structural properties of our experience and certain structural properties of our cognitive processes: the distinctions within phenomenology seem to parallel the distinctions made by our perceptual system, and so on. If we take this principle for granted, we can perhaps "model" the structure of consciousness indirectly by modeling the structure of our cognitive processing: so when Lloyd finds a certain superpositional

structure in a network, this suggests itself as a substrate for the superpositional structure of our experience. This is a worthy explanatory achievement, even though it leaves the key question unanswered: why is there experience at all? These "models" take the existence of experience as a brute fact. Indeed, to gain their explanatory purchase, they must assume wholesale some such principle of structural coherence between experience and processing. But surely, these assumptions are precisely what we want a theory of experience to ultimately explain. Perhaps the existence of experience and these principles of coherence will ultimately need to be taken as basic; but if so, what we will be left with is irreducibility and perhaps even a kind of dualism, rather than the kind of reductive explanation that Lloyd is searching for.

The value of these models, I think, is that they reduce the substrate of processing that underlies conscious experience to its most basic core. Once we get to that core, we find the explanatory gulf remains as wide as ever. Some will be tempted to respond by increasing the complexity of the models, but that ultimately misses the point. The problem with those models is not their simplicity - rather, it is the simplicity of the models that brings out the problems with any models of this kind. At the end of the day, a more complex model of processing will give us just more of the same. A processing model may yield a terrific explanation of functions, abilities, and capacities, but to explain *experience*, it is simply the wrong sort of thing. The moral that I draw is that for a theory of consciousness, we must look elsewhere. But that is a topic for another day.

Pick a Number between Zero and Infinity...

[David J. Chalmers](#)

From: dave@cogsci.indiana.edu (David Chalmers)
Newsgroups: sci.math,sci.math.num-analysis
Subject: Re: call for votes: most & least boring numbers
Date: 17 Jan 90 20:40:02 GMT

In article <18311.25b44848@merrimack.edu> ain14924@merrimack.edu writes:

Reminds me of a friend of mine who claims that the number 17 is "the most random" number. His proof ran as follows: pick a number. It's not really as good a random number as 17, is it?
(Invariable Answer: "Umm, well, no...")

This reminds me of a little experiment I did a couple of years ago. I stood on a busy street-corner in Oxford, and asked passers-by to "name a random number between zero and infinity." I was wondering what this "random" distribution would look like.

The results: (most common numbers first, out of about 150 responses in all):

- 3 (11 people)
- 7 (9 people)
- 5 (8 people)
- 12 (6 people)
- 1, 4, 10, 77 (5 people each)
- 2, 47, infinity-1 (4 people each)
- 15, 17, 20, 27 (3 people each)
- 18, 23, 26, 30, 42, 99 (2 people each)
- 6, 13, 14, 19, 21, 22, 25, thirteen more 2-digit numbers, twenty 3-digit numbers, twelve 4-digit numbers, one 5-digit number, one 6-digit number, four 7-digit numbers, one 8-digit number, one non-integer (328.39), one huge number ($9.265 \cdot 10^{10^{10}}$). (1 person each)

Of course a uniform distribution is a priori impossible so I couldn't have expected that :-). Even a logarithmic distribution is impossible (it has infinite integral). Interestingly enough, this distribution, taken coarsely, was quite close to logarithmic up to 1000 or so. There were roughly the same number of 2-digit responses as 1-digit responses, and a few less 3-digit responses. Then things fell off sharply, however.

Other interesting features:

- 17 wasn't quite as "random" as might have been predicted.
- Extreme frequency of the digit "7" all round.
- Especially notable are the good performances of 77 and 47.
- Poor performance of digit "8", also "6" and "9".
- Both "very prime" (e.g. 17) and "very composite" (e.g. 12) numbers did well.

Then I could tell you about the "random word" experiment I did on Sydney harbour...perhaps another time.

Is the Continuum Hypothesis True, False, or Neither?

[David J. Chalmers](#)

Newsgroups: sci.math

From: chalmers@bronze.ucs.indiana.edu (David Chalmers)

Subject: Continuum Hypothesis - Summary

Date: Wed, 13 Mar 91 21:29:47 GMT

Thanks to all the people who responded to my enquiry about the status of the Continuum Hypothesis. This is a really fascinating subject, which I could waste far too much time on. The following is a summary of some aspects of the feeling I got for the problems. This will be old-hat to set theorists, and no doubt there are a couple of embarrassing misunderstandings, but it might be of some interest to non-professionals.

A basic reference is Gödel's "What is Cantor's Continuum Problem?", from 1947 with a 1963 supplement, reprinted in Benacerraf and Putnam's collection *Philosophy of Mathematics*. This outlines Gödel's generally anti-CH views, giving some "implausible" consequences of CH.

"I believe that adding up all that has been said one has good reason to suspect that the role of the continuum problem in set theory will be to lead to the discovery of new axioms which will make it possible to disprove Cantor's conjecture."

At one stage he believed he had a proof that $C = \aleph_2$ from some new axioms, but this turned out to be fallacious. (See Ellentuck, "Gödel's Square Axioms for the Continuum", *Mathematische Annalen* 1975.)

Two people pointed me to a recent paper by Maddy: "Believing the Axioms", *Journal of Symbolic Logic* 1988 (in 2 parts). This is an extremely interesting paper and a lot of fun to read. A bonus is that it gives a non-set-theorist who knows the basics a good feeling for a lot of issues in contemporary set theory.

Most of the first part is devoted to "plausible arguments" for or against CH: how it stands relative to both other possible axioms and to various set-theoretic "rules of thumb". One gets the feeling that the weight of the arguments is against CH, although Maddy says that many "younger members" of the set-theoretic community are becoming more sympathetic to CH than their elders. There's far too much here for me to be able to go into it in much detail.

Some highlights from Maddy's discussion, also incorporating a few things that other people sent me:

- (1) Cantor's reasons for believing CH aren't all that persuasive today.
- (2) Gödel's proof of the consistency of CH shows that CH follows from ZFC plus the Axiom of Constructibility ($V=L$, roughly that the set-theoretic universe = the constructible universe). However, most set-theorists seem to find Constructibility implausible and much too restrictive. It's an example of a "minimizing" principle, which tends to cut down on the number of sets admitted to one's universe. Apparently "maximizing" principles meet with much more sympathy from set theorists. Such principles are more compatible with $\sim$ CH than with CH.
- (3) If GCH is true, this implies that $\aleph_0$ has certain unique properties: e.g. that it's that cardinal before which GCH is false and after which it is true. Some would like to believe that the set-theoretic universe is more "uniform" (homogeneous) than that, without this kind of singular occurrence. Such a "uniformity" principle tends to imply $\sim$ GCH.
- (4) Most of those who disbelieve CH think that the continuum is likely to have very large cardinality, rather than $\aleph_2$ (as Gödel seems to have suggested). Even Cohen, a professed formalist, argues that the power set operation is a strong operation that should yield sets much larger than those reached quickly by stepping forward through the ordinals:

"This point of view regards C as an incredibly rich set given to us by a bold new axiom, which can never be approached by any piecemeal process of construction."
- (5) There are also a few arguments in favour of CH, e.g. there's an argument that not-CH is restrictive (in the sense of (2) above). Also, CH is much easier to force (Cohen's method) than $\sim$ CH. And CH is much more likely to settle various outstanding results than is $\sim$ CH, which tends to be neutral on these results.
- (6) Most large cardinal axioms (asserting the existence of cardinals with various properties of hugeness: these are usually derived either from considering the hugeness of $\aleph_0$ compared to the finite cardinals and applying uniformity, or from considering the hugeness of V (the set-theoretic universe) relative to all sets and applying "reflection") don't seem to settle CH one way or the other.
- (7) Various other axioms have some bearing. Axioms of determinacy restrict the class of sets of reals that might be counterexamples to CH. Various forcing axioms which are "maximality" principles (in the sense of (2) above), imply $\sim$ CH. The strongest (Martin's maximum) implies that $C = \aleph_2$. Of course the "truth" or otherwise of all these axioms is controversial.
- (8) Freiling's principle about "throwing darts at the real line" is a seemingly very plausible principle, not involving large cardinals at all, from which $\sim$ CH immediately follows. Freiling's paper (JSL 1986) is a good read. More on this at the end of this message.

Of course I've conspicuously avoided saying anything about whether it's even reasonable to suppose that CH has a determinate truth-value. Formalists will argue that we may choose to make it come out

whichever way we want, depending on the system we work in. On the other hand, the mere fact of its independence from ZFC shouldn't immediately lead us to this conclusion - this would be assigning ZFC a privileged status which it hasn't necessarily earned. Indeed, Maddy points out that various axioms within ZFC (notably the Axiom of Choice, and also Replacement) were adopted for extrinsic reasons (e.g. "usefulness") as well as for "intrinsic" reasons (e.g. "intuitiveness"). Further axioms, from which CH might be settled, might well be adopted for such reasons.

One set-theorist correspondent said that set-theorists themselves are very loathe to talk about "truth" or "falsity" of such claims. (They're prepared to concede that $2+2=4$ is true, but as soon as you move beyond the integers trouble starts. e.g. most were wary even of suggesting that the Riemann Hypothesis necessarily has a determinate truth-value.) On the other hand, Maddy's contemporaries discussed in her paper seemed quite happy to speculate about the "truth" or "falsity" of CH.

Personally, not only do I see the integers as bedrock, but I'm also prepared to take any finite number of power sets before I have any problems. At least that far, I'm a diehard Platonist. I'm considerably less sure whether to be a diehard Platonist about the paradise of ridiculously big cardinals. But my intuition is happy with reals and sets of reals, so as a naive non-set-theorist I'm sticking to the belief that CH is determinate one way or the other. As one correspondent suggested, the question of the determinateness of CH is perhaps the single best way to separate the Platonists from the formalists.

And is it true or false? Well, I'd always found CH somewhat intuitively plausible. But after reading all this, it does seem that the weight of evidence tend to point the other way. On the other hand, that's only a second-order approximation based on limited knowledge - more subtle and sophisticated arguments (third-order approximations) might begin to push things the other way. Still, if I had to lay money on it (God, of course, knows the right answer and could settle the bet), I'd feel safer with the money on $\sim$ CH.

I've enclosed (with permission) a brief but enlightening message from Bill Allen on Freiling's Axiom of Symmetry. This is a good one to run your intuitions by.

Many thanks to Bill Allen, James Cummings, David Feldman, Torkel Franzen, Calvin Ostrum, Keith Ramsay, and Peter Suber. I'd be very interested to hear more about this subject.

--Dave Chalmers.

Date: Fri, 8 Mar 91 01:11:37 -0800

From: William C. Allen (allen@math.ucla.edu)

To: chalmers@bronze.ucs.indiana.edu

Subject: Re: Status of the Continuum Hypothesis

Organization: UCLA Mathematics Department

I have a colleague, Chris Freiling, here at UCLA, who, a few years ago proved that a certain plausible assertion is equivalent to the negation of continuum hypothesis (CH) over ZFC. The idea is something like this:

Let A be the set of functions mapping Real Numbers into countable sets of Real Numbers. Given a function f in A , and some arbitrary real numbers x and y , we see that x is in $f(y)$ with probability 0, i.e. x is not in $f(y)$ with probability 1. Similarly, y is not in $f(x)$ with probability 1. Let AX be the axiom which states

"for every f in A , there exist x and y such that x is not in $f(y)$ and y is not in $f(x)$ "

The intuitive justification for AX is that we can find the x and y by choosing them at random.

In ZFC, $AX = \text{not CH}$.

proof:

If CH holds, then well-order R as $r_0, r_1, \dots, r_x, \dots$ with $x < \aleph_1$. Define $f(r_x)$ as $\{r_y : y \leq x\}$. Then f is a function which witnesses the falsity of AX .

If CH fails, then let f be some member of A . Let Y be a subset of R of cardinality $\aleph_1$. Then Y is a proper subset. Let X be the union of all the sets $f(y)$ with y in Y , together with Y . Then, as X is an $\aleph_1$ union of countable sets, together with a single $\aleph_1$ size set Y , the cardinality of X is also $\aleph_1$, so X is not all of R . Let a be in $R \setminus X$, so that a is not in $f(y)$ for any y in Y . Since $f(a)$ is countable, there has to be some b in Y such that b is not in $f(a)$. Thus we have shown that there must exist a and b such that a is not in $f(b)$ and b is not in $f(a)$. So AX holds. -|

I like Freiling's proof, since it does not invoke large cardinals or intense infinitary combinatorics to make the point that CH implies counter-intuitive propositions. Freiling has also pointed out that the natural extension of AX is AXL (notation mine), where AXL is AX with the notion of countable replaced by Lebesgue Measure zero. Freiling has established some interesting Fubini-type theorems using AXL .

See "Axioms of Symmetry: Throwing Darts at the Real Line", by Freiling, Journal of Symbolic Logic, 51, pages 190-200. An extension of this work appears in "Some properties of large filters", by Freiling and Payne, in the JSL, LIII, pages 1027-1035, but Chris tells me he's not as fond of the latter paper as he is the former.

--Bill Allen

--UCLA

From: DJVELLEMAN@amherst.BITNET

Date: Thu, 28 Mar 91 15:03 EST

Subject: Continuum Hypothesis

I just came across your posting about CH and found it quite interesting. A few comments:

One set-theorist correspondent said that set-theorists themselves are very loathe to talk about "truth" or "falsity" of such claims. (They're prepared to concede that $2+2=4$ is true, but as soon as you move beyond the integers trouble starts. e.g. most were wary even of suggesting that the Riemann Hypothesis necessarily has a determinate truth-value.) On the other hand, Maddy's contemporaries discussed in her paper seemed quite happy to speculate about the "truth" or "falsity" of CH.

Many set theorists are loathe to talk about CH being "true" or "false", but I'm not sure how many would agree that trouble starts as soon as you move beyond the integers. I recently did a small survey (only about 10 people) at a set theory conference. I had a list of 10 statements, and for each one you had to choose among 5 choices:

- True
- False
- Either true or false but I don't know which
- Neither true nor false
- Either (either true or false but I don't know which) or (neither true nor false) but I don't know which.

The statements were:

- $2+2=4$
- Fermat's last theorem
- Peano's axioms are consistent
- ZF is consistent
- Fundamental Theorem of Calculus
- Riemann Hypothesis
- AC
- CH
- There are uncountable limit cardinals
- There are measurable cardinals.

One or two people were Platonists the whole way-- restricted their responses to the first three possibilities. Many were Platonists through the Riemann Hypothesis, and then switched to "neither true nor false". I was one of the few who became uncertain once I got beyond the integers. (Remarkably, one person turned in an answer sheet with many answers filled in and crossed out, and wrote "can't decide".

In the end, the only thing he was willing to commit himself to was that $2+2=4$ is true!)

Personally, not only do I see the integers as bedrock, but I'm also prepared to take any finite number of power sets before I have any problems. At least that far, I'm a diehard Platonist. I'm considerably less sure whether to be a diehard Platonist about the paradise of ridiculously big cardinals. But my intuition is happy with reals and sets of reals, so as a naive non-set-theorist I'm sticking to the belief that CH is determinate one way or the other. As one correspondent suggested, the question of the determinateness of CH is perhaps the single best way to separate the Platonists from the formalists.

Suppose we let X_0 =the set of natural numbers, and let X_{n+1} =the power set of X_n . Then you seem to be willing to accept all of these sets as "existing" in some determinate sense. What about their union? What about the power set of their union? If you won't accept these, aren't you drawing a pretty arbitrary line? If you do accept them, then where do you stop? As for using the determinateness of CH as a test to separate Platonists from formalists: I think it's important to recognize that Platonism isn't an all-or-nothing affair. This may be a good test for a certain level of Platonism. I fail--I don't think CH has a determinate truth value. But some people (like intuitionists) don't even think Fermat's last theorem has a determinate truth value. This might be a good test for a lower level of Platonism, and I pass this test--I can't conceive of Fermat's theorem not having a determinate truth value. (By the way, on my survey everyone said Fermat's theorem was either true or false but they didn't know which, except for the guy who only would commit himself to $2+2=4$.)

Dan Velleman
Dept. of Mathematics Amherst College
djvelleman@amherst.bitnet

-
- For much more information, see Nancy McGough's [Continuum Hypothesis Page](#).
 - Return to [Miscellaneous writings](#).
 - Return to [David Chalmers' home page](#).

Thoughts on Emergence

[David J. Chalmers](#)

From: dave@cogsci.indiana.edu (David Chalmers)

Newsgroups: comp.ai.philosophy

Subject: Thoughts on emergence

Date: 6 Oct 90 07:08:47 GMT

Here are some thoughts on "emergence". Nothing definitive, but an attempt to get at the psychological core (or cores) of the notion. Thanks are due to others for providing a stimulating discussion.

Emergence is a tricky concept. It's easy to slide it down a slippery slope, and turn it into something implausible and easily dismissable. But it's not easy to delineate the interesting middle ground in between. Two unsatisfactory definitions of emergence, at either end of the spectrum:

(1) Emergence as "inexplicable" and "magical". This would cover high-level properties of a system that are simply not deducible from its low-level properties, no matter how sophisticated the deduction. This view leads easily into mysticism, and there is not the slightest evidence for it (except, perhaps, in the difficult case of consciousness, but let's leave that aside for now). All material properties seem to follow from low-level physical properties. Very few sophisticated people since the 19th century have actually believed in this kind of "emergence", and it's rarely what is referred to by those who invoke the term favourably. But if you mention "emergence", someone inevitably interprets you as meaning this, causing no end of confusion.

(2) Emergence as the existence of properties of a system that are not possessed by any of its parts. This, of course, is so ubiquitous a phenomenon that it's not deeply interesting. Under this definition, file cabinets and decks of cards (not to mention XOR gates) have plenty of emergent properties - so this is surely not what we mean.

The challenge, then, is to delineate a concept of emergence that falls between the deeply implausible (1) and the overly general (2). After all, serious people do like do use the term, and they think they mean something interesting by it. It probably will help to focus on a few core examples of "emergence":

- (A) **The game of Life:** High-level patterns and structure emerge from simple low-level rules.
- (B) **Connectionist networks:** High-level "cognitive" behaviour emerges from simple interactions between dumb threshold logic units.
- (C) **The operating system** (Hofstadter's example): The fact that overloading occurs just around when there are 35 users on the system seems to be an emergent property of the system.
- (D) **Evolution:** Intelligence and many other interesting properties emerge over the course of evolution by genetic recombination, mutation and natural selection.

Note that in all these cases, the "emergent" properties are in fact deducible (perhaps with great difficulty) from the low-level properties (perhaps in conjunction with knowledge of initial conditions), so a more sophisticated concept than (1) is required. Another stab at a definition might be:

(3) Emergent = "deducible but not reducible". Biological and psychological laws and properties are frequently said not to be reducible to physical laws and properties. For many reasons, not the least being that the high-level laws/properties in question might be found associated with all kinds of different physical laws/properties as substrates. (A universe without protons and electrons might nevertheless include learning and memory.)

There are some problems with this definition, though. Firstly, it's not clear what is gained by trying to explicate emergence in terms of the almost-equally-murky concept of "reduction". Secondly, it seems to let in some not-paradigmatically-emergent phenomena, and it's not clear how some emergent phenomena like (A) or (C) would fit this definition. I think that (3) picks out a very interesting class, but it's not quite the class we're after. It's on the right track, though, I think.

The notion of reduction is intimately tied to the *ease of understanding* one level in terms of another. Emergent properties are usually properties that are more easily understood in their own right than in terms of properties at a lower level. This suggests an important observation: *Emergence is a psychological property*. It is not a metaphysical absolute. Properties are classed as "emergent" based at least in part on (1) the interestingness to a given observer of the high-level property at hand; and (2) the difficulty of an observer's deducing the high-level property from low-level properties. The properties of XOR are an obvious consequence of the properties of its parts. Emergent properties aren't. Might as well give this a number:

(4) Emergent high-level properties are *interesting, non-obvious consequences* of low-level properties.

This still can't be the full story, though. Every high-level physical property is a consequence of low-level properties, usually non-obviously. It feels unsatisfactory, for instance, to say that computations performed by a COBOL program are an emergent property relative to the low-level circuit operations - at least this feels much less "emergent" than a connectionist network. So something is missing. The trouble seems to lie with the complex, kludgy *organization* of the COBOL circuits. The low-level stuff may be simple enough, but all the complexity of the high-level behaviour is due to the complex *structure* that is given to the low-level mechanisms (by programming). Whereas in the case of connectionism or the game of life it feels that we have simplicity in both low-level mechanisms and their organization. So in those cases, we have much more of a "something for nothing" feel. Let's try for another number:

(5) Emergence is the phenomenon wherein complex, interesting high-level function is produced as a result of combining simple low-level mechanisms in simple ways.

I think this is much closer to a good definition of emergence. Note that COBOL programs, and many biological systems, are excluded by the requirement that not only the mechanisms but their principles of

combination be simple. (Of course simplicity, complexity and interestingness are psychological concepts, at least for now, though we might try to explicate them in terms of Chaitin-Kolmogorov-Solomonoff complexity if we felt like it. My intuition is that this is likely to prove a little simplistic, although Chaitin has an interesting paper that attempts to derive a notion of the "organization" of a system using similar considerations.) And note also that most things that satisfy this definition should also satisfy (4) - due to our feeling that simple principles should have simple consequences (or else complex but uninteresting consequences, like random noise). Any complex, interesting consequence is likely to be non-obvious.

This does indeed fit in with the feeling that emergence is a "something for nothing" phenomenon - though in a more subtle and satisfactory way than set forth in (1), for instance. It's a phenomenon whereby "something stupid buys you something smart". And most of our examples fit. The game of Life and connectionist networks are obvious: interesting high-level behaviour as a consequence of simple dynamic rules for low-level cell dynamics. In evolution, the genetic mechanisms are very simple, but the results are very complex. (Note that there is a small difference, in that in the latter case the emergence is diachronic, i.e. over time, whereas in the first two cases the emergence is synchronic, i.e. not over time but over levels present at a given time.)

We're still not completely there - it's not clear how (C), the operating system example, fits into this paradigm of emergence. But throwing in a smidgen of teleology should get us the rest of the way. I.e., we have to notice that everything here has to be relativized to *design*. So we *design* the game of Life according to certain simple principles, but complex, interesting properties leap out and *surprise* us. Similarly for the connectionist network - we only design it at a low level (though in this case we *hope* that complex high-level properties will emerge). Whereas in the COBOL case - and in the case of much traditional AI - you only get out what you put in (N.B. I'm not necessarily knocking this: at least here, I'm trying to explicate emergence, not to defend it). And now the operating system example fits in well. The design principles of the system in this case are quite complex - unlike the other cases that fit (5) above - but still the figure "35" is not a part of that design at all. So:

(6) Emergence is the phenomenon wherein a system is designed according to certain principles, but interesting properties arise that are not included in the goals of the designer.

Notice the appearance of the word "goal" - this is important, any design is goal-relative. So the notion now is quite teleological. I notice that Russ Abbott makes a similar point in a recent posting. Notice, however, that as we've conceded that emergence is a psychological property, we're able to construe teleology in a psychological, non-absolute way. So for our purposes here, we only need the *appearance* of teleology. This is nice, because it allows us to include system where strictly speaking, "design" doesn't apply at all. In evolution, for instance, there is no "designer", but it is easy to treat evolutionary processes as processes of design. On more than one level.

We can view evolution as teleological at the level of the gene - as in Dawkins' theory, for instance. Then the appearance of complex, interesting high-level properties such as intelligence is quite emergent. We also can reconstrue evolution as teleological at the level of the organism (this is perhaps a more

straightforward Darwinian view of things). On this construal, the most salient adaptive phenomena like intelligence are no longer emergent, but the goal of the design process. However, this view does open up the possibility of other kinds of emergent phenomena: firstly, non-selected-for byproducts of the evolutionary process (such as Gould and Lewontin's "Spandrels"); secondly and more intriguingly, it allows an explanation for why "consciousness" (or "subjectivity" or "qualia" or whatever) seems emergent. Raw consciousness doesn't not seem to have been selected for, as it doesn't play any direct functional role (though it does have functional counterparts; this is a subtle issue, but remember we're talking about the way things *seem*, not the way they are); but it somehow emerges as a byproduct of selection for adaptive process such as intelligence.

It's probably foolish to search for a definitive construal of "emergence": like most psychological concepts, it probably is best construed as a "family resemblance" - each of the "definitions" outlined above might play some role. Personally, I'm happiest with a combination of (5) and (6) - with (5) being the "core" variety of emergence, and (6) being a more general variety of which (5) is a special case.

From: dave@cogsci.indiana.edu (David Chalmers)

Newsgroups: comp.ai.philosophy

Subject: Re: Thoughts on emergence

Date: 8 Oct 90 22:56:22 GMT

In article <1990Oct7.012523.3828@watdragon.waterloo.edu> cpshelley@violet.uwaterloo.ca (cameron shelley) writes:

While intelligence might not be considered emergent at the "Darwinian" level synchronically, it must still be so diachronically. Unlike a 'designer', natural selection can only passively enforce changes and so must wait until the random process of mutation gives it the proper opportunity. Therefore, at the time when a significant enough increase in intelligence occurs, it must be emergent. So intelligence should still be considered emergent in the diachronic sense you mentioned earlier (and I cut out :).

It depends on how you construe the teleology of the evolutionary process. Of course there's no real teleology there, but psychologically we can construe it in at least three different ways. My original post suggested two of these ways, your response suggests a third.

First way: Teleology at the level of the gene. Goal: get genes to replicate. At this level, intelligence is emergent.

Second way: Teleology at the level of the organism. Goal: get organisms to survive and replicate. This is your suggestion. If the only teleology we impute to the evolutionary process is that of a "natural selector" - i.e. produce systems that survive and replicate - then specific functions like intelligence are indeed emergent.

Third way: Teleology at the level of the organism. Goal: get organisms to be like X (for some X, e.g. X = strong, fast, intelligent...). Here, we are construing the teleology of the evolutionary process not as that of the natural selector, but as that of the "blind watchmaker". On this view, intelligence is the goal of the evolutionary process, and so cannot be regarded as emergent.

The teleology here not being metaphysical but psychological, all these ways of construing it are quite valid. My original post only mentioned the first and third possibilities. The second possibility is also a very reasonable construal, and serves the useful purpose of showing how intelligence can be regarded as emergent without having to descend to the level of the "selfish gene". Sorry for omitting this possibility in my original post.

Realms of Cognitive Science

[David J. Chalmers](#)

From: chalmers@bronze.ucs.indiana.edu (David Chalmers)
Subject: Re: Two Materialisms (Was: Re: Strong AI and Panpsychism)
Date: Thu, 30 Jan 92 03:50:35 GMT

B Chandrasekaran writes:

It appears that there are three realms: the realm of matter, the realm of representations, and the realm of qualia/intentions/consciousness, not just two: matter and consciousness.

I like this distinction, although I think there might more naturally be four realms to distinguish.

- 1. **Physical realm:** atoms, quantum physics, gravitation, electromagnetism, ...
- 2. **Functional realm:** feedback loops, networks, Turing machines, ..
- 3. **Representational realm:** knowledge, belief, desire,..
- 4. **Consciousness realm:** consciousness, qualia, pains, sensations, ...

On the relations between the various realms:

The functional realm is uncontroversially reducible to the physical realm.

Modern cognitive science holds that the representational realm is reducible to the functional realm, and I agree, although some, e.g. Searle, disagree, holding that it is only reducible to the physical realm, and no doubt you could find some who don't accept even that (e.g. people who hold that representation is dependent on consciousness, and that consciousness isn't reducible to the physical).

Consciousness is where things get tricky. Some people think that consciousness is reducible to representation (e.g. Dennett, Harman, and possibly many cognitive scientists, though cognitive scientists don't talk much about consciousness). Many more think that consciousness is reducible to function (Lewis, Armstrong, McDermott, most functionalists, etc), but not necessarily to representation. Still others (notably Searle) thinks that consciousness is reducible to the physical, but not to the other realms. Incidentally Searle also differs from most others in holding that representation is itself dependent on consciousness.

Penrose may in some strange sense hold that consciousness is reducible to function, but believes that the right function can only be realized through quantum effects (but he's so vague that one can't know). A Cartesian dualist like Eccles doesn't think that consciousness is reducible to physics, but might just possibly think that it's reducible to function that happens to be realized by ghostly spirits.

Then there are people like me who can't see how consciousness could be reduced to physics (and consequently, to function or representation either), though it would be nice if somehow it could be.

Finally there are people who don't think that consciousness really exists as a phenomenon worth talking about (perhaps Minsky, and Dennett on some days), so there's no point reducing it to anything. Personally I think that apart from the strong anti-reductionist position, this is the only position that's at all tenable, although it does seem to be *prima facie* implausible.

In my paper on connectionism in AI Magazine a few years ago, I argued that part of the excitement surrounding connectionism in its early stages was the claim of skipping the second realm, i.e., it was more directly physical and not representational. Most connectionists quickly abandoned this notion, since they came to see that it was every bit as representational as the symbolic systems, only different. Edelman seems to think he is offering a purely physicalist account, without any representational intervention, but I think there is representation there nevertheless.

I think that the enlightened connectionist position has always been that the cognitive realm (which is more or less the representational realm) is reducible to the functional realm - you don't find many connectionists talking directly in terms of physics. It's just that they want to reduce it to functional mechanisms quite different from those used by traditional AI and cognitive science.

To be strictly accurate I should probably introduce the cognitive realm as a fifth realm - i.e. the realm concerned with the explanation of behaviour - and not beg any questions about the identification of this with the representational realm. Some cognitive scientists want to skip the representational realm altogether, and go straight from the cognitive to the functional (e.g. Brooks, Freeman, the Churchlands on some days, and possibly some of the connectionists you mention). Most cognitive scientists, by contrast, take it for granted that cognitive explanation will be representational explanation; or more accurately, jointly representational and functional explanation. A few people (Fodor, Pylyshyn, ...) might hold that one can perform the reduction to the representational realm fully before a reduction to the functional realm, but even these people will rely on nonrepresentational functional explanations for some cognitive phenomena, e.g. perception. But I'd better stop now. The space of realms is getting cluttered.

TIME Magazine

March 25, 1996 Volume 147, No. 13

Return to [Contents page](#)

CAN MACHINES THINK?

Maybe so, as Deep Blue's chess prowess suggests. and that sparks a fresh debate about the nature of mind. Is it just neurons?

ROBERT WRIGHT

When Garry Kasparov faced off against an IBM computer in last month's celebrated chess match, he wasn't just after more fame and money. By his own account, the world chess champion was playing for you, me, the whole human species. He was trying, as he put it shortly before the match, to "help defend our dignity."

Nice of him to offer. But if human dignity has much to do with chess mastery, then most of us are so abject that not even Kasparov can save us. If we must vest the honor of our species in some quintessentially human feat and then defy a machine to perform it, shouldn't it be something the average human can do? Play a mediocre game of Trivial Pursuit, say? (Or lose to Kasparov in chess?)

Apparently not. As Kasparov suspected, his duel with Deep Blue indeed became an icon in musings on the meaning and dignity of human life. While the world monitored his narrow escape from a historic defeat--and at the same time marked the 50th birthday of the first real computer, ENIAC--he seemed to personify some kind of identity crisis that computers have induced in our species.

Maybe such a crisis is in order. It isn't just that as these machines get more powerful they do more jobs once done only by people, from financial analysis to secretarial work to world-class chess playing. It's that, in the process, they seem to underscore the generally dispiriting drift of scientific inquiry. First Copernicus said we're not the center of the universe. Then Darwin said we're just protozoans with a long list of add-ons--mere "survival machines," as modern Darwinians put it. And machines don't have souls, right? Certainly Deep Blue hasn't mentioned having one. The better these seemingly soulless machines get at doing things people do, the more plausible it seems that we could be soulless machines too.

But however logical this downbeat argument may sound, it doesn't appear to be prevailing among scholars who ponder such issues for a living. That isn't to say philosophers are suddenly resurrecting the

idea of a distinct, immaterial soul that governs the body for a lifetime and then drifts off to its reward. They're philosophers, not theologians. When talking about some conceivably nonphysical property of human beings, they talk not about "souls" but about "consciousness" and "mind." The point is simply that as the information age advances and computers get brainier, philosophers are taking the ethereal existence of mind, of consciousness, more seriously, not less. And one result is to leave the theologically inclined more room for spiritual speculation.

"The mystery grows more acute," says philosopher David Chalmers, whose book *The Conscious Mind* will be published next month by Oxford University Press. "The more we think about computers, the more we realize how strange consciousness is."

Though chess has lately been the best-publicized measure of a machine's humanity, it is not the standard gauge. That was invented by the great British computer scientist Alan Turing in a 1950 essay in the journal *Mind*. Turing set out to address the question "Can machines think?" and proposed what is now called the Turing test. Suppose an interrogator is communicating by keyboard with a series of entities that are concealed from view. Some entities are people, some are computers, and the interrogator has to guess which is which. To the extent that a computer fools interrogators, it can be said to think.

At least that's the way the meaning of the Turing test is usually put. In truth, midway through his famous essay, Turing wrote, "The original question, 'Can machines think?,' I believe to be too meaningless to deserve discussion." His test wasn't supposed to answer this murky question but to replace it. Still, he did add, "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."

Guess again. With the century's end in sight, no machine has consistently passed the Turing test. And on those few occasions when interrogators have been fooled by computers, the transcripts reveal a less-than-penetrating interrogation. (Hence one problem with the Turing test: Is it measuring the thinking power of the machines or of the humans?)

The lesson here--now dogma among researchers in artificial intelligence, or AI--is that the hardest thing for computers is the "simple" stuff. Sure they can play great chess, a game of mechanical rules and finite options. But making small talk--or, indeed, playing Trivial Pursuit--is another matter. So too with recognizing a face or recognizing a joke. As Marvin Minsky of the Massachusetts Institute of Technology likes to say, the biggest challenge is giving machines common sense. To pass the Turing test, you need some of that.

Besides, judging by the hubbub over the Kasparov match, even if computers could pass the test, debate would still rage over whether they think. No one doubted Deep Blue's chess skills, but many doubted whether it is a thinking machine. It uses "brute force"--zillions of trivial calculations, rather than a few strokes of strategic Big Think. ("You don't invite forklifts to weight-lifting competitions," an organizer of exclusively human chess tournaments said about the idea of man-vs.-machine matches.) On the other

hand, there are chess programs that work somewhat like humans. They size up the state of play and reason strategically from there. And though they aren't good enough to beat Kasparov, they're good enough to leave the average Homo sapiens writhing in humiliation.

Further, much of the progress made lately on the difficult "simple" problems--like recognizing faces--has come via parallel computers, which mirror the diffuse data-processing architecture of the brain. Though progress in AI hasn't matched the high hopes of its founders, the field is making computers more like us, not just in what they do but in how they do it--more like us on the inside.

So machines can think? Not so fast. Many people would still say no. When they talk about what's inside a human being, they mean way inside--not just the neuronal data flow corresponding to our thoughts and feelings but the thoughts and feelings themselves. You know: the exhilaration of insight or the dull anxiety of doubt. When Kasparov lost Game 1, he was gloomy. Could Deep Blue ever feel deeply blue? Does a face-recognition program have the experience of recognizing a face? Can computers--even computers whose data flow precisely mimics human data flow--actually have subjective experience? This is the question of consciousness or mind. The lights are on, but is anyone home?

For years AI researchers have tossed around the question of whether computers might be sentient. But since they often did so in casual late-night conversations, and sometimes in an altered state of consciousness, their speculations weren't hailed as major contributions to Western thought. However, as computers keep evolving, more philosophers are taking the issue of computer consciousness seriously. And some of them--such as Chalmers, a professor of philosophy at the University of California at Santa Cruz--are using it to argue that consciousness is a deeper puzzle than many philosophers have realized.

Chalmers' forthcoming book is already making a stir. His argument has been labeled "a major misdirection of attention, an illusion generator," by the well-known philosopher Daniel Dennett of Tufts University. Dennett believes consciousness is no longer a mystery. Sure there are details to work out, but the puzzle has been reduced to "a set of manageable problems."

The roots of the debate between Chalmers and Dennett--the debate over how mysterious mind is or isn't--lie in the work of Dennett's mentor at Oxford University, Gilbert Ryle. In 1949 Ryle published a landmark book called *The Concept of Mind*. It resoundingly dismissed the idea of a human soul--a "ghost in the machine," as Ryle derisively put it--as a hangover from prescientific thought. Ryle's juiciest target was the sort of soul imagined back in the 17th century by Rene Descartes: an immaterial, somewhat autonomous soul that steers the body through life. But the book subdued enthusiasm for even less supernatural versions of a soul: mind, consciousness, subjective experience.

Some adherents of the "materialist" line that Ryle helped spread insisted that these things don't even exist. Others said they exist but consist simply of the brain. And by this they didn't just mean that consciousness is produced by the brain the way steam is produced by a steam engine. They meant that the mind is the brain--the machine itself, period.

Some laypeople (like me, for example) have trouble seeing the difference between these two views--between saying consciousness doesn't exist and saying it is nothing more than the brain. In any event, both versions of strict materialism put a damper on cosmic speculation. As strict materialism became more mainstream, many philosophers talked as if the mind-body problem was no great problem. Consciousness became almost passe.

Ryle's book was published three years after ENIAC's birth, and at first glance his ideas would seem to draw strength from the computer age. That, at any rate, is the line Dennett takes in defending his teacher's school of thought. Dennett notes that AI is progressing, creating smart machines that process data somewhat the way human beings do. As this trend continues, he believes, it will become clearer that we're all machines, that Ryle's strict materialism was basically on target, that the mind-body problem is in principle solved. The title of Dennett's 1991 book says it all: *Consciousness Explained*.

Dennett's book got rave reviews and has sold well, 100,000 copies to date. But among philosophers the reaction was mixed. The can-do attitude that was common in the decades after Ryle wrote--the belief that consciousness is readily "explained"--has waned. "Most people in the field now take the problem far more seriously," says Rutgers University philosopher Colin McGinn, author of *The Problem of Consciousness*. By acting as if consciousness is no great mystery, says McGinn, "Dennett's fighting a rearguard action."

McGinn and Chalmers are among the philosophers who have been called the New Mysterians because they think consciousness is, well, mysterious. McGinn goes so far as to say it will always remain so. For human beings to try to grasp how subjective experience arises from matter, he says, "is like slugs trying to do Freudian psychoanalysis. They just don't have the conceptual equipment."

Actually there have long been a few mysterians insisting that the glory of human experience defies scientific dissection. But the current debate is different. The New Mysterians are fundamentally scientific in outlook. They don't begin by doubting the audacious premises of AI. O.K., they say, maybe it is possible--in principle, at least--to build an electronic machine that can do everything a human brain can do. They just think people like Dennett misunderstand the import of such a prospect: rather than bury old puzzles about consciousness, it resurrects them in clearer form than ever.

Consider, says Chalmers, the robot named Cog, being developed at M.I.T.'s artificial-intelligence lab with input from Dennett (see following story). Cog will someday have "skin"--a synthetic membrane sensitive to contact. Upon touching an object, the skin will send a data packet to the "brain." The brain may then instruct the robot to recoil from the object, depending on whether the object could damage the robot. When human beings recoil from things, they too are under the influence of data packets. If you touch something that's dangerously hot, the appropriate electrical impulses go from hand to brain, which then sends impulses instructing the hand to recoil. In that sense, Cog is a good model of human data processing, just the kind of machine that Dennett believes helps "explain" consciousness.

But wait a second. Human beings have, in addition to the physical data flow representing the heat, one

other thing: a feeling of heat and pain, subjective experience, consciousness. Why do they? According to Chalmers, studying Cog doesn't answer that question but deepens it. For the moral of Cog's story seems to be that you don't, in principle, need pain to function like a human being. After all, the reflexive withdrawal of Cog's hand is entirely explicable in terms of physical data flow, electrons coercing Cog into recoiling. There's no apparent role for subjective experience. So why do human beings have it?

Of course, it's always possible that Cog does have a kind of consciousness--a consideration that neither Dennett nor Chalmers rules out. But even then the mystery would persist, for you could still account for all the behavior by talking about physical processes, without ever mentioning feelings. And so too with humans. This, says Chalmers, is the mystery of the "extraness" of consciousness. And it is crystallized, not resolved, by advances in artificial intelligence. Because however human machines become--however deftly they someday pass the Turing test, however precisely their data flow mirrors the brain's data flow--everything they do will be explicable in strictly physical terms. And that will suggest with ever greater force that human consciousness is itself somehow "extra."

Chalmers remarks, "It seems God could have created the world physically exactly like this one, atom for atom, but with no consciousness at all. And it would have worked just as well. But our universe isn't like that. Our universe has consciousness." For some reason, God chose "to do more work" in order "to put consciousness in."

When Chalmers says "God," he doesn't mean--you know--God. He's speaking as a philosopher, using the term as a proxy for whoever, whatever (if anyone, anything) is responsible for the nature of the universe. Still, though he isn't personally inclined to religious speculation, he can see how people who grasp the extraness of consciousness might carry it in that direction.

After all, consciousness--the existence of pleasure and pain, love and grief--is a fairly central source of life's meaning. For it to have been thrown into the fabric of the universe as a freebie would suggest to some people that the thrower wanted to impart significance.

It's always possible that consciousness isn't extra, that it actually does something in the physical world, like influence behavior. Indeed, as a common-sense intuition, this strikes many people as obvious. But as a philosophical doctrine it is radical, for it would seem to carry us back toward Descartes, toward the idea that "soul stuff" helps govern the physical world. And within both philosophy and science, Descartes is dead or, at best, on life support. And the New Mysterians, a pretty hard-nosed group, have no interest in reviving him.

The extraness problem is what Chalmers calls one of the "hard" questions of consciousness. What Dennett does, Chalmers says, is skip the "hard" questions and focus on the "easy" questions--and then title his book *Consciousness Explained*. There is one other "hard" question that Chalmers emphasizes. It--and Dennett's alleged tendency to avoid such questions--is illustrated by something called pandemonium, an AI model that Dennett favors.

According to the model, our brain subconsciously generates competing theories about the world, and only the "winning" theory becomes part of consciousness. Is that a nearby fly or a distant airplane on the edge of your vision? Is that a baby crying or a cat meowing? By the time we become aware of such images and sounds, these debates have usually been resolved via a winner-take-all struggle. The winning theory--the one that best matches the data--has wrested control of our neurons and thus of our perceptual field.

As a scientific model, pandemonium has virtues. First, it works; you can run the model successfully on a computer. Second, it works best on massively parallel computers, whose structure resembles the brain's structure. So it's a plausible theory of data flow in the human brain, and of the criteria by which the brain admits some data, but not other data, to consciousness.

Still, says Chalmers, once we know which kinds of data become part of consciousness, and how they earned that privilege, the question remains, "How do data become part of consciousness?" Suppose that the physical information representing the "baby crying" hypothesis has carried the day and vanquished the information representing the rival "cat meowing" hypothesis. How exactly--by what physical or metaphysical alchemy--is the physical information transformed into the subjective experience of hearing a baby cry? As McGinn puts the question, "How does the brain 'turn the water into wine?'"

McGinn doesn't mean that subjective experience is literally a miracle. He considers himself a materialist, if in a "thin" sense. He presumes there is some physical explanation for subjective experience, even though he doubts that the human brain--or mind, or whatever--can ever grasp it. Nevertheless, McGinn doesn't laugh at people who take the water-into-wine metaphor more literally. "I think in a way it's legitimate to take the mystery of consciousness and convert it into a theological system. I don't do that myself, but I think in a sense it's more rational than strict materialism, because it respects the data." That is, it respects the lack of data, the yawning and perhaps eternal gap in scientific understanding.

These two "hard" questions about consciousness--the extraneous question and the water-into-wine question--don't depend on artificial intelligence. They could occur (and have occurred) to people who simply take the mind-as-machine idea seriously and ponder its implications. But the actual construction of a robot like Cog, or of a pandemonium machine, makes the hard questions more vivid. Materialist dismissals of the mind-body problem may seem forceful on paper, but, says McGinn, "you start to see the limits of a concept once it gets realized." With AI, the tenets of strict materialism are being realized--and found, by some at least, incapable of explaining certain parts of human experience. Namely, the experience part.

Dennett has answers to these critiques. As for the extraneous problem, the question of what function consciousness serves: if you're a strict materialist and believe "the mind is the brain," then consciousness must have a function. After all, the brain has a function, and consciousness is the brain. Similarly, turning the water into wine seems a less acute problem if the wine is water.

To people who don't share Dennett's philosophical intuitions, these arguments may seem unintelligible. (It's one thing to say feelings are generated by the brain, which Chalmers and McGinn believe, but what

does it even mean to say feelings are the brain?) Still, that doesn't mean Dennett is wrong. Some people share his intuitions and find the thinking of his critics opaque. Consciousness is one of those questions so deep that frequently people with different views don't just fail to convince one another, they fail even to communicate. The unintelligibility is often mutual.

Chalmers isn't a hard-core mysterian like McGinn. He thinks a solution to the consciousness puzzle is possible. But he thinks it will require recognizing that consciousness is something "over and above the physical" and then building a theory some might call metaphysical. This word has long been out of vogue in philosophy, and even Chalmers uses it only under duress, since it makes people think of crystals and Shirley MacLaine. He prefers "psychophysical."

In *The Conscious Mind*, Chalmers speculatively sets out a psychophysical theory. Maybe, he says, consciousness is a "nonphysical" property of the universe vaguely comparable to physical properties like mass or space or time. And maybe, by some law of the universe, consciousness accompanies certain configurations of information, such as brains. Maybe information, though composed of ordinary matter, is a special incarnation of matter and has two sides--the physical and the experiential. (Insert Twilight Zone music here.)

In this view, Cog may indeed have consciousness. So might a pandemonium machine. So might a thermostat. Chalmers thinks it quite possible that AI research may someday generate--may now be generating--new spheres of consciousness unsensed by the rest of us. Strange as it may seem, the prospect that we are creating a new species of sentient life is now being taken seriously in philosophy.

Though Turing generally shied away from such metaphysical questions, his 1950 paper did touch briefly on this issue. Some people, he noted, might complain that to create true thinking machines would be to create souls, and thus exercise powers reserved for God. Turing disagreed. "In attempting to construct such machines we should not be irreverently usurping his power of creating souls, any more than we are in the procreation of children," Turing wrote. "Rather we are, in either case, instruments of his will providing mansions for the souls that he creates."

The New York Times

[Home](#)[Sections](#)[Contents](#)[Search](#)[Forums](#)[Help](#)

April 16, 1996

Arizona Conference Grapples With Mysteries of Human Consciousness

By SANDRA BLAKESLEE

TUCSON, Ariz. -- Like the proverbial blind men trying to identify by sense of touch a large, thick-hided animal with a trunk at one end and a tail at the other, some of the world's top scientists, philosophers and far-out thinkers gathered here last week to contribute their different perspectives on the elephant of consciousness.

A good time was had by all, even when the fur -- or maybe it was elephant hide -- began to fly.

Can machines be conscious? The question elicited a spirited debate between those who said, Of course, it's just a matter of time and clever engineering, and others who replied: Never! It's bad enough that you think consciousness can arise from gray lumps of tissue. It is inconceivable that sentience could ever emerge from wholly insentient matter.

Then there were less contentious questions. Does free will exist? Can consciousness exist without emotions? Are animals conscious? What happens to your conscious mind when you fall into a deep sleep?

And the most debated question of all: is consciousness something very special and unique or is it just the natural byproduct of a complex brain, emerging like wind from intricate weather patterns?

The conference, "Toward a Science of Consciousness," was sponsored by the University of Arizona with support from the Fetzer Institute and the

Institute of Noetic Sciences, two organizations dedicated to exploring the metaphysical foundations of Western science. A similar but smaller conference was held two years ago in Tucson.

"We were deliberately eclectic in choosing speakers," said Dr. Stuart Hameroff, an anesthesiologist at the Arizona Health Sciences Center who was a principal organizer of the event. Experts in various disciplines can always talk among themselves, he said, "but there needs to be an arena where everyone can mix their ideas about consciousness together."

Thus the conference drew neuroscientists, philosophers, mathematicians, computer scientists, physicists, dream researchers, pharmacologists, doctors, ethnologists, psychologists, parapsychologists, scholars of religion and a variety of prophets who claim to have solved the mystery of consciousness.

The meeting was unusual from the start. Dr. Jaron Lanier, a computer scientist from Columbia University who is a pioneer in virtual reality, opened the plenary session on Monday by playing a brief piano recital. His blond dreadlocks flew apace with the music. The audience was delighted.

The goal of the meeting was simple, Hameroff said. What is the nature of consciousness? Can we hope to understand it scientifically?

It is remarkable that such a diverse gathering could discuss the question of consciousness in a coherent manner. But this kind of cross pollination of ideas, where everything goes, is exactly what is needed, said Dr. Christof Koch, a neuroscientist at the California Institute of Technology who helped organize the event. One hundred years ago, people could not understand how life could arise out of mere chemicals, he said. But when DNA was explained, theories of vitalism -- that a magical force was needed to explain life -- disappeared.

The study of consciousness is like the study of physics before Newton, said Dr. Piet Hut, a theoretical physicist at the Institute for Advanced Studies in Princeton, N.J. In fact, he said, if people had organized a conference about physics in the Middle Ages, they would have dismissed Copernicus and Galileo as crackpots. "We shouldn't make that mistake today," Hut said.

But before progress can be made on the question, some definitions are in order. Consciousness has many guises.

In Tucson, the tone of discourse was set by a young philosopher from the

University of California at Santa Cruz, Dr. David Chalmers. He is widely credited for posing the so-called hard problem of consciousness.

To explain this concept, Chalmers first described the so-called easy problems of consciousness, the sorts of questions being tackled in neuroscience laboratories around the world: How does sensory information get integrated in the brain? How do we see and reach out for an object? How are we able to verbalize our internal states and report what we are doing or feeling?

"These problems are not trivial," Chalmers said. "They may take 100 years or more to solve, but progress is being made."

The hard problem is this: What is the nature of subjective experience? Why do we have vividly felt experiences of the world? Why is there someone home inside our heads?

Thus far, nothing in physics or chemistry or biology can explain these subjective feelings, Chalmers said. "What really happens when you see the deep red of a sunset or hear the haunting sound of a distant oboe, feel the agony of intense pain, the sparkle of happiness or meditative quality of a moment lost in thought?" he asked. "It is these phenomena, often called qualia, that pose the deep mystery of consciousness."

In Tucson, people mounted four responses to the hard problem: it doesn't exist, it will be answered soon enough by conventional science, there must be something else in the universe that we do not yet understand, and hey guys, forget it, we can never understand consciousness.

Dr. Daniel Dennett, a philosopher at Tufts University, is a forceful proponent of the idea that consciousness is no big deal. "It's like fame," Dennett said. "It doesn't exist except in the eye of the beholder." When does fame happen, he asked? Is it when 10 people know your name? A hundred people? A thousand?

Scientists have shown that information coming into the brain is broken down into separate processing streams, Dennett said. But no one has yet found any "place" where all the information comes together, presenting a whole picture of what is being felt or seen or experienced. The temptation, he said, is to believe that the information is transduced by consciousness. But it is entirely possible that the brain's networks can assume all the roles of an inner boss. Mental contents become conscious by winning a competition against other

mental contents, Dennett said. No more is needed. Consciousness is an epiphenomenon.

A second group of scientists agreed with Dennett but took a softer line. When all the "easy" problems are solved, the hard problem will disappear -- but consciousness certainly exists. "It's silly to deny it," said Dr. Pat Churchland, a philosopher at the University of California at San Diego.

Awareness and subjectivity are network effects involving many millions of nerve cells in the cortex and thalamus, Dr. Churchland said. And while the exact nature of the phenomenon cannot yet be explained, the call for a "new physics" or some mysterious forces in nature are not needed.

Dr. Rodolfo Llinas, a neuroscientist at New York University, agreed, suggesting that timing effects inside the brain produce conscious experience.

Those who believe machines can someday be conscious tended to fall into this camp. The trick will be to make computers that are sufficiently complex, said Dr. Danny Hillis, vice president of research and development at Walt Disney Imagineering in Glendale, Calif. Then, like human brains, they should give rise to the emergent properties of consciousness.

Others tried to answer a few of the easy questions. Dr. Allan Hobson, a sleep expert at Harvard Medical School, described a neurobiological theory of dreaming. It does not explain where consciousness "goes" when people are asleep, he said, but finds that different chemical states in the brain seem to produce different sorts of consciousness.

The next major group of consciousness seekers might be called modern dualists. Agreeing with the hard problem, they feel that something else is needed to explain people's subjective experiences. And they have lots of ideas about what this might be.

According to Chalmers, scientists need to come up with new fundamental laws of nature. Physicists postulate that certain properties -- gravity, space-time, electromagnetism -- are basic to any understanding of the universe, he said.

"My approach is to think of conscious experience itself as a fundamental property of the universe," he said. Thus the world has two kinds of information, one physical, one experiential. The challenge is to make theoretical connections between physical processes and conscious

experience, Chalmers said.

Another form of dualism involves the mysteries of quantum mechanics. Dr. Roger Penrose from the University of Oxford in England argued that consciousness is the link between the quantum world, in which a single object can exist in two places at the same time, and the so-called classical world of familiar objects where this cannot happen.

Moreover, with Hameroff, he has proposed a theory that the switch from quantum to classical states occurs inside certain proteins called microtubules. The brain's microtubules, they argue, are ideally situated to perform this transformation, producing "occasions of experience" that with the flow of time give rise to stream of consciousness thought.

The notion came under vigorous attack. "Pixie dust in the synapses is about as explanatorily powerful as quantum mechanics in the microtubules," Churchland said. Their logic is, consciousness is deeply mysterious, quantum mechanics is deeply mysterious, ergo the two are the same mystery, she said.

Penrose's ideas are popular, Churchland said, because many people have dualist hankerings. They want to believe in a soul, life after death and the specialness of humans and their inner thoughts. They have a negative gut reaction to the idea that neurons -- cells that can be probed under a microscope -- are the source of the "me-ness of me," she said.

Finally, there are those who argued that people can never understand consciousness. The mystery is too deep. Dr. Colin McGinn, a philosopher from Rutgers University, said that for humans to grasp how subjective experience arises from matter "is like slugs trying to do Freudian psychoanalysis -- they just don't have the conceptual equipment."

But this did not deter many from trying. During the week, presentations were made on animal consciousness (featuring apes, dolphins and gray parrots), free will and the spiritual nature of consciousness.

Dr. Robert Forman, a professor of religion at Hunter College in New York, said mystical experience had something to tell people about consciousness. "To understand genes," he said, "we look at bacteria like E. coli. To study memory, we analyze the memory of a sea slug. But to probe consciousness, we need to examine the experience of mystics, who experience their own consciousness in its simplest form."

Millions of people regard these types of experiences, feeling a oneness with the universe, as the highest experience that the conscious brain has to offer, Forman said.

Copyright 1996 The New York Times Company

How to Make a Soul

By Andrew Brown

Among the professionals, philosophical argument is a martial art. Nervous spectators pull their beer bottles out of the way in case a backhand logic chop should sweep them off the table. This is personal.

The young man defending himself in the Empire Bar is dressed like a jobbing rock star: t-shirt, tight jeans around skinny legs, great curly masses of brown hair reaching down his back. But he moves like a fighter; every point launched against him is blocked, deftly, with his outspread palms.

His opponent is older, still wiry, with black hair and moustache. The more the young man blocks, the harder the older man's arms throw new points. Soon he is arguing from his shoulders, like a boxer. Finally, he seizes a beer bottle and thrusts it in front of his opponent. "Look," he shouts. "There's only one thing I want to know. Do you think this beer bottle has consciousness?"

There's a pause. "Well, it might have," says David Chalmers. A ripple of appreciative relaxation runs round the audience. The bout is over. One of the spectators takes the bottle from Bruce Mangan's hand and carefully tears a strip from the label. He waves it in front of Chalmers. "So what happens now?" he asks. "Where has the consciousness gone in the paper? Has the label got its own little consciousness?"

It's a game. It is also extremely serious. The crowd around Chalmers and Mangan at the Empire Bar is part of a congregation of neuroscientists, philosophers, psychologists, quantum physicists and AI gurus that has come in pilgrimage to Tucson for a conference called [Towards a Science of Consciousness](#). There are almost a thousand people attending, and each participant worth talking to has at least half a dozen good theories, mostly incompatible. What holds all the theories and disciplines together is simple. These people think that the nature of consciousness is the most exciting intellectual frontier in the world today. It is here that science seems to be closing in on the essence of what makes us human.

Yet there is a paradox at the heart of this work. The more the researchers learn, the further the true subject recedes. The more is learned about how the brain works, the less this knowledge can be tied up with how it feels to have a brain - to feel oneself living and thinking. Science has always discussed the world from the outside. Can it ever encompass

how the world feels from the inside?

There is no doubt that the process of learning from the outside is getting better and better. You can use a brain scanner to look inside someone's head while he learns a poem or recites one back. You can see which parts of the brain are working, and how hard. But even if these scanners could track each signal in each of the brain's billions of cells, they would still not explain why that activity should feel the way it does. People may, someday, be able to find which pattern of neurons firing makes up a thought. But what does that tell you about consciousness, about the feeling of the thought?

Chalmers calls the discovery of the patterns in the brain that correspond to thinking the easy problem. By calling it easy, he does not mean to deny that it is also fiendishly difficult. It would mark the culmination of centuries of work in neurology, in bio-chemistry, in experimental psychology. Answer that easy problem and you have the mechanics of memory, learning, sensation laid out before you.

But easy is a relative term. To Chalmers, such questions are easy because people have some idea of how, in principle, they might be answered. This contrasts with the question of why consciousness arises from a physical process. That "why" is what Chalmers calls the hard problem. As he outlines its hardness, he speaks in a light, hurried voice. His words come in quick flurries, like snow.

Watching Chalmers talk, you realise you can see less and less; you're stuck deeper and deeper in a world of improbable contradictions. "When it comes to the hard problem, my feeling is that you need something that goes beyond physical theory, because everything in physical theory is compatible with the absence of consciousness; my feeling is that you have to take consciousness as axiomatic, like time and space. "The problem comes with constructing a theory that will link them all together. You want to get down to something that's deep and fundamental - a set of laws that's simple enough you can write them on the front of a t-shirt." The scientific study of consciousness, Chalmers says, "is like physics before Isaac Newton came along. No one knows what is really happening."

A new way of looking at the world, expressible in some practical, useful axioms. A call for a new Newton. No wonder that the brilliant and ambitious flocked to the Tucson conference, the second of its ilk and the foremost gathering of its type in the world, to try their mettle.

All you Zombies

Go back a few decades and you will be hard put to find any inkling of a science of consciousness. Behaviourists treated animals and humans as black boxes; neurologists

were still puzzled by the synapse. Even the study of perception was academically suspect. But fashions change, and new tools become available, and those two things shape all science. When Francis Crick, one of the discoverers of the structure of DNA, announced in the '80s that he was going to spend the rest of his life studying consciousness, he was responding to, and amplifying, a growing sense that consciousness would be the next big thing.

The idea that some animals are conscious is now almost as common among the scientists who long rejected it as it always has been with everyone else. "Most people have no difficulty in seeing consciousness in cats or dogs, though once you get down to flies it's more difficult." says Chalmers. "But there may be some very simple form of consciousness, experience, without much in the way of thought or activity - something about consciousness that is pre-intellectual." Then, as an afterthought: "Some of our machines may have that now."

With this remark he leaps across one of the great chasms that divide the field. Nothing has done more to sharpen the issues involved in consciousness research than the promise, or the spectre, of artificial intelligence (AI). As Chalmers, who studied for two years under AI guru [Douglas Hofstadter](#), puts it, "The deep question is why any physical system, whether machine or animal, is associated with consciousness. But brains did it, so why shouldn't machines, too?" In other words, when the hard question is solved, artificial intelligence in its truest form is thrown in as an added bonus. The problem of making a computer that knows itself to be capable of thought, one with an inner life, joins the ranks of the easy questions.

Some people think it already belongs there. Dan Dennett is a large, long-legged man, with a great rounded skull like an ostrich egg, and a beard like God (in whom he does not believe). Wherever he stepped in the corridors or anterooms of the Tucson conference he became the immediate focus of a ragged, admiring ellipse of students and disputants upon whom he beamed with sharp benevolence. He is the inventor of one of the classical thought experiments of AI: the replacement of a grey porridgy brain - yours, for the sake of argument - with a shiny new one made of something utterly different.

Cell by cell, bit by bit, your neurons would be turned to silicon. But each piece of silicon would have exactly the same connections and behaviour as the cell it replaced. The neighbouring cells would get exactly the same responses as they would if the delicate electric and chemical feelers of their dendrites were still brushing against other cells, rather than just being plugged into the wiring. Would you notice? Would you care? And if you did, why?

No one has yet found anything magic about the way the neurons in the brain signal to one another. It is immensely complicated, but it's only chemicals and electricity. It can be

measured and mimicked. In the outer suburbs of the brain it already has been. There is a treatment for deafness which replaces a defective nerve with circuitry. Treatments for blindness that would replace the eye with a television camera are already thinkable. Why not replace a whole brain, and thus show that silicon can support consciousness as well as carbon does?

This sort of reasoning by equivalence has a long pedigree in AI. If the discipline has a founding principle, it is that of the Turing test: the idea that a computer which could converse as convincingly as a human could be said to have artificially captured the human intellect. By taking a similar line of thought rather further, Dennett concludes that the hard problem, as defined by Chalmers, is no more than a mirage. When all the "easy" problems about how the brain processes information have been solved, we will discover that the hard problem has simply disappeared.

Indeed, one of the things that makes the hard problem so difficult at the moment is that lots of people can't see that it is a problem at all. People like Dennett - believers in "Strong AI", also known, for obvious reasons, as zombies - are sure that consciousness will turn out to be no more than the sum of meaningless mathematical manipulations of information.

Dennett defends these views with enormous clarity, force and charm. Nonetheless, he seems to be drawing back in recent years from the particularly strong position with which he made his name. Though he still says that, in principle, any chunk of silicon (or anything else) that can perform the same functions as a brain would by definition be conscious, he now admits that some of these functions seem to be much more specialised now than they did even ten years ago. The more we know about the human brain, the less it seems likely that we can ever reproduce anything like it artificially. Among those who still believe it can be done, the methods which were fashionable when Dennett first started work are long forgotten. Danny Hillis, for example, the founder of Thinking Machines, one of the first successful parallel processing companies, and now vice-president of R&D for Disney, believes that the only way to get a machine complicated enough to have a possibility of consciousness is to breed it:

"Imagine something like the Internet, multiplied times 100, and imagine all the machines on it exchanging programs, and imagine using those programs to design a system which would run not on one machine but on the whole network - then I think you have the image of something that might be complicated enough to be conscious," Hillis told the conference. "Once you do that, it becomes easier to accept the idea of something like a conscious machine. I think people who have a strong intuition that machines can't be conscious have that feeling not because they overestimate the wonders of consciousness but because they underestimate the powers of machinery. A lot of arguments against machines thinking are made from exaggeration and distortion."

Hillis's foremost target is Sir Roger Penrose, a brilliant Oxford mathematician who is partly responsible for the conference's being in Tucson. Penrose believes that consciousness is fundamentally not like the operations of a computer, and cannot be recreated by means of those operations. Rather, he believes it has to be tied to a new set of natural laws, laws that combine the apparent randomness of the quantum world with Einstein's general theory of relativity. And laws that would explain the fundamental sympathy for mathematical form that Penrose, among other things a brilliant geometer, sees in the human mind.

After putting this set of ideas into a demanding, beautifully written and, say the strong AI types, deeply wrong-headed book - *The Emperor's New Mind* - Penrose came into contact with another non-traditional consciousness researcher, Stuart Hameroff. Hameroff's fascination with consciousness comes from his professional need to extinguish it; he is an anaesthesiologist at the University of Arizona's hospital. And he's fascinated by the fact that consciousness, whatever it may be, can be removed and restored with a subtle mix of simple gasses.

Hameroff came up with the idea that microtubules, tiny pipes that stiffen the insides of cells, were crucial to the whole story. Penrose got interested and the two began to collaborate. The first Tucson conference, two years ago, was one of the fruits of that collaboration.

The proteins that form the walls of the microtubules can flip between two different shapes. They can also exist, for an instant or two, in a state of quantum superposition, like Schroedinger's cat; it is the flickering through that third state that Penrose and Hameroff now believe constitutes conscious events. Their theory is resonant, especially when Hameroff explains that psychedelic drugs promote superposition in the microtubules, or when he adds that the consciousness event is a blister in space-time, and so must be quickly reduced to normal physics. "If not reduced, a blister in space-time would shear off into multiple universes - and we hate it when that happens." If ambition were all, the Penrose-Hameroff theory and its implication that consciousness is built into the very fabric of nature would score highly. However, most people reckon its chances of success are small enough to slip between a couple of electron shells. And even if the microtubules function as they are supposed to in this theory, the difficulty pointed out by Patricia Churchland, a philosopher, remains: that there is no clear account of how these quantum events should cause consciousness. The hard problem remains untouched.

Cogito Ergo Something

You don't have to believe Penrose and Hameroff to disagree with the strong AI view that a machine could be conscious. When Hillis put the question to his audience, the majority professed itself agnostic: the 60% expressing a view was pretty evenly split between pros

and antis. But it has to be said that the antis had all the best tunes. Literally so, in [Jaron Lanier's](#) case. He opened the proceedings with a self-composed piano piece that started off strange and brilliant, and as it went on became progressively less strange. This formed a counterpoint to his reasoning, which was unflaggingly brilliant but grew stranger and stranger as he argued that talk of machine consciousness veers between the futile and the really dangerous.

As well as being a musician, Lanier is a code god: he invented the term "virtual reality", and then built the gadgetry to make it real. He has a towering physical presence, being well over six feet tall and about four feet round the waist, with a great shock of maize-coloured dreadlocks reaching almost to his waist and a beard that gives his face a perfectly triangular jaw like an Egyptian painting's. His eyes are a bright, pale blue.

He's a genius, of course, and knows it in the same sort of way he knows he's tall. Talking to him, I felt as if I was talking to the spirit of the prairies: a restless, inexhaustible wind running on forever. America, he says, loves frontiers, and always peoples them with abstractions, like freedom, or computing. But it is all a mistake. The abstractions do not exist; the frontiers are only a trick of perspective.

First he disposes of the Turing test. Only a fucked-up gay Englishman being tortured with hormone injections could possibly have supposed that consciousness was some kind of social exam you had to pass, he says. There are two logically possible ways to for a machine to pass the Turing test. Either it gets smarter, or we get dumber. He is in no doubt that the likelier reason would be that we have got stupider and adapted our reasoning to the machine's expectations.

Then he demolishes the idea of computation. It is an arbitrary process, he says. It is not built into the structure of the universe, but only into the way we understand it. Anything can be measured in numbers, and any set of numbers can constitute a program for some thinkable computer. Even a meteor shower could be read as numbers, and so be read as a program on some computer somewhere. Would we therefore say that the meteor shower was computing, or conscious?

Then he has a go at the physical reality of the universe. We were sitting outside as he said this, facing each other across a concrete table which was one of the most solid objects you could hope to find. But, he insisted, on the level of quantum electrodynamics (QED), the most precisely confirmed of all physical theories, it turns out to be an illusion.

"I hate to tell you this, but QED does not acknowledge the existence of gross objects. From the point of view of physics, there are particles in positions, but no gross objects; the particles don't constitute a table." He gestures at the stone table where we sit in the dusty evening, scents of sage-brush and diesel borne round us on a gentle wind. "Gross objects

aren't needed by physics. They're actually extraneous. Ask a neutrino."

I feel as if the prairie wind had whisked me off to Oz. "This is genuinely hard stuff," he says. "Hard to think about and hard to talk about." Then he returns to the more mundane objections to a mechanistic view of consciousness. It narrows our vision of the world. The danger of believing that consciousness is computation is not that it causes us to misunderstand machines, but that it makes us misunderstand our own nature.

Lanier places the hopes that some people have for consciousness research firmly in the American tradition of faith in some apocalyptic fix to all life's problems. Indeed, the programmatic atheism of the main conference was thrown into high relief by the simple faith of some of the attendees that technology could perform all the functions once expected of God. This irrational longing gives the field of consciousness research a lot of its excitement, yet it is only the very naive who will admit to it.

I had supper one night with a man who had made enough money in mainframes never to need to work again. He could travel the world gratifying his curiosity - and the miracle he expected from consciousness research was immortality. Humans, he felt, should be migrated onto better hardware when the old stuff wears out. This is the natural development of the classic Dennett thought experiment where you replace each part of my brain by a bit of silicon, thought by byte, until I wake up one day to find myself running on entirely different hardware, as if I, a mere human, were as clever and well designed as System 7.5.

Love and DNA

That was far from the strangest thing I heard, even if it was one of the more poignant. There were over 500 papers delivered at Tucson, and many of them were deliciously cranky. One Californian paper displayed in a poster session, where anyone could pin their ideas up on a notice-board, claimed that "Pioneering scientific research has demonstrated the experience of love to facilitate coherent harmonic patterns in the EKG, to improve both immune function and hormonal regulation, and enable focused attention to modify the conformation of human DNA. Love has also been shown to be causal in modulating the structure of water to provide a matrix for information storage and transformation."

If love doesn't suit you as an explanation, try evolution. Physiologists and philosophers argue about what consciousness is, but there is also a raft of questions about why it is, what function it might fulfil. Here the analytical tools come easily to hand. If the question is how did a biological phenomenon come about, the answer has to include evolution. Consciousness needs to pay its evolutionary way.

It might seem obvious that an animal that deals with the world in an intelligent, coordinated fashion has an advantage. But that does not translate into a need for consciousness. Most of the intelligent, coordinated decisions crucial for an animal's survival are made far faster than the conscious mind can process things. Jeffrey Gray of the [Institute of Psychiatry](#) in Denmark Hill, South London, serves up an example from Centre Court. In the time that it takes a tennis player at Wimbledon to become consciously aware that his opponent's serve has crossed the net towards him, he must already have struck his own return ball if he is to stay in the game. Even quite high-level trains of action, like commuting by car, or talking over breakfast with your spouse, can be performed without any conscious input. So why should consciousness be selected by evolution?

Gray's answer is that consciousness works rather like a newspaper: it does not report everything that happens, only the things that are unexpected and unpredictable. In a loop that takes about half a second to complete, he says, he believes "the contents of consciousness perform a monitoring function to check whether the actual and predicted states of the perceptual world match or not." And this continuous activity is where we live.

The other function widely ascribed to consciousness is that of a global workspace. Bernard Baars, a sturdy, bearded Californian, tanned like a hazelnut, claims that consciousness takes place in an area of the brain accessible to all the separate sub-processes of our mind, any one of which may suddenly come to dominate the public arena. The technical details, as with Gray's theory, are hugely complicated. But all these theories, along with ten or twenty others, are still attempts at the easy problems. The fact that consciousness is being approached as a field of scientific study will inevitably change the way that people think about themselves, irrespective of the hard scientific results. The science provides new metaphors. They take on a life of their own. The consequences of people thinking of themselves as soft machines may matter more, in some ways, than the question of whether or not they are soft machines. Take free will, as Colin Blakemore discussed in Tucson. Blakemore, an English perceptual physiologist whose experiments on the eyesight of new-born kittens have made him something of a cause celebre among animal-rights campaigners, argues that free will must be an illusion. What happens in the mind is conditioned by what happens in the brain; what happens in the brain is determined by physical laws. Therefore, if you know the state of the brain at any one moment, you can predict its next state, and its next, and so on. Free will is then exposed as an illusion. If only we knew enough, we could see that all our states of mind are equally conditioned by the brain's physical, law-bound realities.

This position looks like science. If it doesn't feel like the truth, then consider a fact revealed by related psychological research: the brain is great at self deception. We are constantly aware of what is going on in our heads, and much of what we are aware of is false. The process of turning the blooming, buzzing confusion of the world into a coherent story involves a tremendous amount of suppression, distortion and illusion. For instance, there is a well-known experiment in which the subject is wired up to a machine that

monitors the brain, given a button to push and told to push it whenever he feels like it - and to say whenever he feels the urge to push the button. That's all. It would seem an untrammelled exercise of free will. Yet the experimenters, watching, know when the subject is going to push the button before he himself is aware that he is about to, or wants to; there is a particular burst of electrical activity in his brain half a second before the subject himself realises that he wants to press the button.

Because of this and similar effects, Blakemore proposes that the distinction between voluntary and involuntary acts is illusory - no more than "folk psychology", to use the jargon sneer. This is a view with potentially enormous consequences. David Hodgson, an Australian judge who has taken a keen interest in the problems of consciousness research, points out that in his world, the distinction between voluntary and involuntary acts is essential.

"I think this raises a really hard question for people who want to do away with folk psychology of voluntary action, belief and intention, and replace it with the concepts of neuroscience, which have a reputable place in the scientific world of cause and effect," Hodgson told the conference. "What do they see as replacing the consent of the woman as the crucial factor in determining whether an act of sexual intercourse is lawful or non-lawful? And if they can't give a non-evasive answer to that question, then I think they should reconsider their programme."

Perhaps the answer could be supplied by the Secret Policeman's Brain Scanner. In the best traditions of philosophy and computer science, this is a device the efficiency of which is unimpaired by the fact that it hasn't actually been built; I invented it half way through the conference. But it is already a very useful tool for examining consciousness researchers. It consists of a small, portable headset full of clever little scanning widgets, and a screen that the experimenter can study. It gives a complete readout of the state of the brain at any one moment, just like a hardware debugger can in silicon. In short, on a personal level, it answers all Chalmers's easy questions. To some, like Dennett and, it turns out, Blakemore, the scanner thus answers everything, providing the policeman with better knowledge of the subject than the subject himself has access to. David Chalmers sees it as less clear cut, showing "the broad outline of what's going on and the likely behavioural dispositions."

Neuropsychologists, though, turned out to be less sure that the secret policeman's brain scanner would be useful. There are two main reasons for the scepticism. The first is that the functional unit of the brain seems not to be the neuron but, instead, networks of neurons, and that these networks are constantly shifting coalitions, with no more stability or definition than a cloud of fireflies. What is more, each neuron can be a member of several different firefly clouds at once, playing different roles in each. Quite precise regions of the brain can be mapped down as essential to certain functions - usually after a small localised stroke has stopped the functioning. But memories and ideas appear to be distributed all over the cortex, linked only by associations that are different in each

individual brain.

Douglas Watt, a neuropsychologist from Quincey, Massachusetts, says that the whole project of the secret policeman's scanner is doomed because our brains are just too complex, too individual. There would be no useful way to interpret the data. Each brain is unique; each has its own genes, its own history, its own consciousness. Even the simplest networks, such as the one in our speech area that recognises a single syllable, may be located a couple of million neurons away in your brain from where it is in mine. Also, the exact location of this syllable switch - not only which neurons are involved but also which other neurons they talk to - will shift over time, particularly if I learn other languages and distinguish more carefully between phonemes. This means that, to be able to use the secret policeman's scanner at a given time, you might have to have been running it throughout the scannee's life. The meaning of a snapshot image would have to be derived from its history; consciousness can only be understood as a narrative.

One does not need a human brain to establish this; the classic experiment uses rats. They were taught to recognise a particular smell, and a wonderfully subtle arrangement of electrodes detected the exact area of the olfactory bulb in their brains which twinkled into life when they did so. So far so good for the secret policeman's scanner; we know that when these particular neurons go off, the rat is conscious of a particular smell. But the next stage lifted the experiment into the realms of genius.

The brain-mapped rats were trained to associate this smell with some other stimulus - a flashing light, or a buzzer, or another of the usual forms of entertainment laid on for laboratory rats. After that training, the rats were once more exposed to the same smell, without the additional stimulus - and this time an entirely different group of neurons fired off. Even to a rat, it seems, there is no such thing as a pure sensation. Everything is stored, and experienced, in relation to other sensations and to emotions. The story of consciousness is one of feelings.

"A recent project at Yale found there was nearly no such thing as an emotionally neutral word," says Watt. "Orienting information supplied by low-level emotion is essential for working memory, and working memory is essential for consciousness. We have two cultural icons, Data and Spock, who are clearly sentient beings, and show better purpose than most of us, and yet claim to have no emotion. I think this is just impossible and these characters are just contradictions in terms."

One-handed clapping

If consciousness is built on feeling, then feelings about feelings may prove particularly fruitful in its study. One way to get access to them, as to so many things, is to look at what

happens when things are disrupted; in the brain, this usually means a stroke. Small strokes can explode in the brain like smart bombs, taking out one function, but no more than that. A blood vessel bursting here will rob you of the power of speech; another in a slightly different place will leave you able to speak and hear, but not to understand, language; another will produce "blindsight", a condition in which the patient can actually see things but does not know that he can, and so considers himself blind.

Vilayanur Ramachandran, of the University of California, San Diego, told the Tucson conference about a particularly odd class of stroke effects. His patients are women who have not only been paralysed down one side by a stroke, but have also been robbed by the calamity of the knowledge that this has happened to them.

If someone "normally" paralysed is asked to pick up a tray of drinks, he will use his one good hand to pick it up from the middle. If one of Dr Ramachandran's patients is asked to do so, she will grasp one side of the tray as if her left hand was grasping the other, and lift confidently. "Oh, how clumsy I am" she will exclaim when she spills the tray's contents everywhere. His patients simply cannot see that one of their hands is not taking part in the process. They are lucid in all other respects: they are able to tell him when and where they had a stroke, but simply unable to admit even to themselves that this stroke has paralysed them.

He describes one patient who was convinced that her left hand, which could not move at all, was touching his nose. "I couldn't resist the temptation ... I said, 'Mrs B: can you clap?' "She said, 'Of course I can clap.' "I said, 'Clap!' "She went" (he moves his right hand in a lurching motion through the air to the point where it would have met the left hand). "This has profound philosophical implications," he continues, as laughter ripples round the conference hall, "because it answers the age-old Zen master's riddle - 'what is the sound of one hand clapping?' You need a damaged brain to answer this question." Dr Ramachandran follows his strange findings into unpopular waters. For decades now, nothing could have been more unfashionable in serious academic psychology than Freud. Yet what Ramachandran sees reminds him inescapably of Freudian theories of denial, repression and other defence mechanisms. He believes that the pattern of denial his stroke patients exhibit points to the mind's continuous struggle to produce a coherent picture of the worlds, and to prefer coherence to accuracy - a very Freudian notion.

In Ramachandran's view, the struggle is between the brain's hemispheres. When isolated facts are reported which might upset the mind's currently held view of the world, the reaction of the left hemisphere is to ignore them. Most of the time, this will be the correct response; sensory systems are not perfect. But the right hemisphere carries out the occasional reality check, just to be sure, and if it thinks something's awry, it gets together with the left hemisphere and, quite literally, changes the mind. In stroke patients who cannot recognise their condition this mechanism stops working. The right hemisphere messages never get through and then, he says, "There is no limit to the delusions that the

left hemisphere will engage in."

The condition is not permanent. Though it will reassert itself, it can be dissipated for a few moments by squirting ice-cold water into the ear on the unparalysed side. The effect is easy to miss, because if you squirt cold water into the wrong ear, as Dr Ramachandran did the first time he tried it, you are left with a patient who is confused, and angry that anyone should have squirted cold water without warning or reason into her ear, but still unaware that she is paralysed. But if the water is squirted into the ear of the damaged hemisphere the patient experiences a period of confusion and then about ten minutes when she knows perfectly well that she has been paralysed - cannot imagine not knowing this, in fact. Six hours later, she will have forgotten the whole episode, and once more be convinced that everything is working properly.

Making a Soul

I know how they feel. The Tucson conference was a series of strange perceptions and iced water in the ear, of epiphanies that slide away, never quite recorded.

On that night in the Empire Bar, after Bruce Mangan and David Chalmers had finished their sparring, [Patrick Wilken](#), one of the spectators, turned to me. He is executive editor of [Psyche](#), the most severe of the three journals of consciousness research to have appeared in the last five years. He also moderates the Psyche-D mailing list, one of the few places on the Internet where real work gets done in public. "You realise what all this is about," he said. "We're trying to invent a new soul."

Over the days that statement grew on me. Instead of deep sleep, I could only dream. Instead of dreaming, I would wake up. As I lay in my hotel room in the thin desert dawn, I couldn't see the sky over the mountains, blue and purple like the flank of a rainbow trout. I could only see little pack trains of cholinergic and aminergic neurons trudging up and down my brain, from its roots to its thoughts. One morning I woke up like this but with a jingle in my head, too: "There's a fire that calls me miracle; that will not cause me chemical" - the phrase repeated over and over, up and down a scale like the trudging neurons. I was eavesdropping on my brain trying to make sense of itself. Whoever I was.

This exalted, jangling state went on until I began my journey home. Outside Dallas, a razor-cut horizon bled fresh dawn into a new hotel room. And after a night of sleep I woke to find that Patrick Wilken's new soul finally made sense. It was not, as I had thought, the soul of a machine we were trying to make; it was our own souls we were inventing here. I fell back into a sleep rich with dreams.

Andrew Brown (abrown@lazy.demon.co.uk) is a brain in a vat. He believes himself to be

the religious affairs correspondent for The Independent.

This article was originally published in August 1996 UK edition of Wired Magazine.

*Copyright 1996 Wired Ventures - info@wired.co.uk
Wired UK, Shand House, 14-20 Shand Street, London SE1 2ES, UK
vox: +44 (0) 171 775 3446; fax: +44 (0) 171 775 3401*

FRONT
PAGESouthern
California
Auto GroupClick Here
Over 23
Franchises

Wednesday, October 16, 1996

[PREV STORY](#)[NEXT STORY](#)
[PDF EDITION](#)
[POLITICS & POLLS](#)[FRONT PAGE](#)[NATION & WORLD](#)[STATE & LOCAL](#)[SPORTS](#)[BUSINESS &
TECHNOLOGY](#)[LIFE & STYLE](#)[CALENDAR](#)[COMMENTARY](#)[WEEKLY SECTIONS](#)[ORANGE COUNTY](#)[SAN FERNANDO
VALLEY](#)[VENTURA COUNTY](#)

COLUMN ONE

Unraveling the Riddle of Identity

■ Exotic devices that allow the brain to be caught in the act of thinking let researchers explore science's last frontier--the source of human consciousness.

By ROBERT [LEE HOTZ](#), Times Science Writer

I have something in mind--a thought so evanescent that it comes and goes in milliseconds.

To capture it in the act, UCLA neuroscientist Mark S. Cohen has trained on my brain a 22-ton experimental imaging device, twice as powerful as--and 30,000 times faster than--any conventional medical imager. This nuclear magnetic resonance (NMR) imager generates a magnetic field so forceful that, as it pulses and flexes around my head, the room shakes with a 100-decibel pile-driver roar.

Every 50 milliseconds, it captures an image of the physical spark of my imagination at work--and something more.

Only part of my conscious mind attends to the mental exercise that Cohen is recording. The rest is abuzz--worried whether I can perform the task properly, curious about how well my mind will photograph, and wondering most of all about the motives of these scientists who are so intent on the neurobiology of my thoughts.

This irrepressible mental chorus constitutes the background noise in the NMR image Cohen takes of my brain. It is also the core of the most perplexing problem in science: What is consciousness?

The human brain, and the self-awareness that arises from it, is a mystery wrapped in an enigma, swaddled in a tough, protective membrane and sealed inside the skull, unknowable until now except by the most indirect means.

Until recently, only a surgeon ever saw an exposed, living brain. A neuroscientist could learn more about its intangible mental

functions by reading poetry or arguing philosophy than by examining the dead organ on an autopsy table.

Today, scientists have at hand an array of exotic devices that can peer through the skull to catch the living brain at work. Researchers at UCLA and other centers are using them to explore one of the last uncharted territories: the structure and cognitive functions of the human brain.

Emboldened by their ability to capture the image of something as intangible as imagination, a growing number of scientists are trying to study the one thing that many believe cannot be labeled, scrutinized or even defined--human consciousness.

The effort to understand human consciousness is an inquiry older than science itself. It is at the heart of a riddle of identity that has preoccupied philosophers, mystics and theologians for as long as there have been words to frame the query: Who--or what--am I?

Many neuroscientists hope that by studying the neurons of the brain, its genes, sensory perceptions, memory and language systems, they may be able to collect enough information about the way it works to finally discover the organizing principles underpinning all subjective experience.

Consequently, researchers for the first time feel confident enough to frame serious questions about the physical foundations of the human spirit.

How do the physical processes of the brain give rise to subjective experience?

Somehow, the fragile synapses and cells of the human nervous system can perceive the world around them, learn from their perceptions, reinforce memories with the force of emotion, plan ahead, decide and act on incomplete information, as well as sleep, dream, wake and pay attention.

The brain captures its moods in melodies. It invents stock markets, founds religions and orbits telescopes. It is introspective enough to develop psychoanalysis.

"There is something very mysterious about consciousness," said Christoff Koch, a theoretical neurobiologist at Caltech. "Why can objective physical systems have subjective states? It is baffling.

"It gets at the central idea of the soul."

For the human brain, the scientific investigation of consciousness is the beginning of an unusual journey of self-discovery.

"We are trying to understand who we are by studying the organ that allows you to understand who you are," said Antonio Damasio, an expert at the University of Iowa on the brain, cognition and

behavior.

"Consciousness," said David Chalmers, a cognitive scientist and philosopher at UC Santa Cruz, "is the last frontier of science."

Chemistry and Electricity

Unraveling the nature of consciousness, however, is a problem too daunting for any one scientist to address in its entirety.

Instead, researchers are teasing apart isolated neural processes, such as how the brain perceives color or how neural cells focus attention, in the hope that they can gather evidence of how a conscious mind is assembled from different brain processes.

On scores of university campuses, scientists trace the ebb and flow of chemicals that trigger brain functions, the blood flows that nourish them and the electrical patterns they generate. They are trying to catch the mind in the act of being.

*

As they pull together new insights into how the brain functions, a growing number of neuroscientists believe that the mental processes underlying consciousness arise from an intricate madrigal of two languages--chemistry and electricity--communicated through networks of millions of neurons, all orchestrated precisely in time.

"We are the activity of the neuronal machine," said Rudolfo Llinas, an expert at New York University on the brain and cognition.

For the first time, researchers are identifying the connections between the brain's physical anatomy and the mechanisms of perception, learning and other, higher cognitive functions.

Investigators are beginning to understand how experiences can lodge permanently among millions of scattered neurons--to be revived in milliseconds by the smell of a baking pastry, a familiar melody or a photograph.

Memory is more complicated than anyone had imagined. Researchers such as Larry Squire at UC San Diego have discovered that the brain harbors many independent systems of memory, while Erin Schumann at Caltech and her colleagues are revising ideas of how knowledge is retained in the chemistry of nerve cells.

Scientists have determined that the brain handles memories of events and emotions differently than memories of ingrained habits and tasks.

Short-term memories are fundamentally different from long-term memories and may be forged in entirely different ways. False memories appear to be handled differently in the brain than memories based on true events, brain scans suggest.

Memory's unexpectedly intricate arrangements mirror those of

other neural systems involved in the operations of a conscious mind.

Until recently, most scientists believed that language was handled by the left side of the brain. But the mechanisms of language are much more decentralized than previously thought. And the way the growing brain handles language is much more flexible than previously believed, research by Elizabeth Bates at UC San Diego suggests.

Names of animals and names of tools are handled by largely separate brain regions, other scientists discovered. Knowledge of the concepts they describe is stored in another discrete system.

*

The words, syntax and concepts of a person's native tongue are stored so separately from any language learned later in life that a stroke can knock out the ability to speak one but not the other.

By highlighting nuances of brain functions, sophisticated biomedical sensors offer glimpses of tantalizing relationships between physical brain structures and conscious mental activity:

- * Subtle differences in brain anatomy appear to affect the ways men and women process information, even when thinking about the same things, hearing the same words or solving similar problems.

- * The most efficient brains appear also to be the smartest, with the brains of those with the highest IQs using the least energy. Learning and practice appear to improve the brain's efficiency.

- * Small structural abnormalities appear to develop in the brains of people with Alzheimer's disease or Huntington's chorea long before any noticeable behavioral symptoms can be diagnosed.

- * Minor alterations in neural circuits for vision and hearing may be responsible for dyslexia, while brain abnormalities in regions involved in inhibiting mental activity could be the cause of attention deficit hyperactivity disorder, which affects about 5% of school-age children.

But the search for clues to consciousness leads down stranger alleys.

From nature's own experiments--the victims of disease and brain injury--neurologists are obtaining glimpses into the workings of the conscious mind by seeing how defects twist its usual attributes.

*

What is one to make of an otherwise normal man--a victim of a rare mental disorder called Charles Bonnet syndrome--who occasionally sees leprechauns, the amputee who thinks he is holding a cup in the hand he no longer has, or the engineer who

sees Technicolor cartoon characters cavorting in a blind spot in her visual field.

Stranger still are victims of a mental disorder called associative agnosia, who easily can draw any object but cannot recognize what it is, or those with "blindsight," whose sense of sight functions normally but who cannot make themselves consciously aware of what their eyes see.

Why should brain damage leave one woman unable to name animals, but able to name any other object?

From these and dozens of other hints, researchers today are identifying the puzzle pieces of the brain, yet they are far from understanding how they fit together into the cohesive whole of the human mind.

"Why does all this processing give you an inner life?" asked UC Santa Cruz's Chalmers. "Why is it that these processes should give rise to the consciousness in the first place?"

Our Uniquely Human Self-Consciousness

Some researchers argue that consciousness in some form may extend beyond humans to other species.

Each human mind may be unique, but all its higher cognitive functions take place in a brain closely resembling those of more primitive primates such as apes and chimpanzees, say experts in neural evolution.

The human brain appears to have no unique cells, chemicals, neural circuits or major anatomical structures, experts said. All the known differences are a matter of degree.

Other species can plan ahead based on experience, combining sensory stimuli, attention and short-term memory. Nor is the human mind unique in its ability to form concepts, some researchers contend. Even pigeons appear to be able to sort objects into categories and recognize abstract relationships.

What seems to distinguish human consciousness is its subjective self-awareness--the feeling of being "me."

And that is where most scientists throw up their hands in frustration, because human self-awareness is itself the best, and perhaps only, evidence of its existence: I know I have a mind because I have a mind that is aware of itself.

"How do you explain this quantum leap in control over the mind in humans, with basically the same machinery you have in monkeys?" asked neurobiologist Martin C. Sereno at UC San Diego. "People just have more control over what is going on in their heads--for better or worse."

*

There is no shortage of new theories from scientists and philosophers trying to account for the novelty of human consciousness.

Some contend that consciousness is the byproduct of a Darwinian tooth-and-claw competition between mental states trying to dominate behavior.

On the other hand, Oxford University mathematician Roger Penrose proposed that consciousness must be a property of abstruse particle physics operating inside brain cells. Free will, he contends, must arise from the random character of quantum mechanics--a rolling of the subatomic dice.

Some experts in artificial intelligence conclude that consciousness is nothing more than sophisticated computer-like information processing. Special mental switching areas integrate all the brain's perceptions into a sense of self, they suggest.

Indeed, if the raw meat of the human brain can develop consciousness, they speculate, why can't the silicon of integrated computer circuits?

Neuroscientists like Sereno and philosophers like Daniel Dennett at Tufts University believe that language is one essential prerequisite for the development of higher consciousness. How, after all, can a brain be self-aware if it cannot talk to itself?

Damasio thinks they have it backward. In his view, language developed well after the onset of awareness. The brain harbored a sense of self long before it had a word to name it, he said.

"If you look at a chimpanzee, I have no doubt it has a sense of self, just not as rich as yours or mine," he said. "What we have is enriched by a spectacular memory of our past and by a memory of the plans of our future, combined with the fact that we automatically translate whatever we think into language.

"That is truly, uniquely human."

However intoxicating the theoretical possibilities, Nobel laureate Francis Crick at the Salk Institute in La Jolla insists on keeping his ideas about consciousness tied as closely as possible to what can be measured in a laboratory.

"There was too much talk and not enough experiments, from my point of view," he said. "You really want new experimental data and to not get into too many highfalutin arguments about the nature of consciousness."

*

In Crick's view, the sweeping question of human consciousness is too large to study. Instead he has teased off a more manageable piece of the problem.

With his collaborator Christoff Koch, Crick is looking for the neural mechanisms underlying visual awareness.

They are considering the role played by special neural cells in the higher visual cortex. Some cells respond to straight lines while other cells react to color, movement or even faces. To assemble a coherent image, rich with all the sensory impressions, memories, knowledge and emotions it conjures, these separate visual cells must somehow link up to all the other neurons that relate to what the eyes are seeing.

At the highest levels of this process, Koch and Crick believe, some cells must respond to the unified perception assembled from so many other active brain cells. It is those special cells, they speculate, that make someone "conscious" of what the brain has perceived.

"People think consciousness must arise from the incredible complexity of millions of interconnected neurons," Koch said.

"We don't think it is pushing the envelope to say there are specific neurons that give rise to consciousness."

Theater of the Mind

The scanner has finished imaging my brain.

My head, still enveloped in an imaging cradle, slides out of the massive machine like a spent artillery shell.

The experiment was deceptively simple.

I had been asked to tap the fingers of my right hand together slowly, then quickly, then more slowly again. Then I was asked to imagine doing the same tasks.

The machine recorded my brain functions by tracking minute variations in the scanner's magnetic field caused by alterations in neural blood flow.

It takes Cohen, who helped design the scanner, hours on a high-speed computer to transform those thousands of individual measurements into a picture of the brain that a human eye can comprehend.

Projected on a color monitor, the scans yield rows of wrinkled raisin-shaped images of my brain. Daubs of bright color highlight the regions that became active when I physically moved my fingers and those regions that became active when I imagined the same action.

Both tasks--real and imagined--lighted up the same areas of the brain responsible for movement.

At some fundamental level, thinking about an action and performing it appear to be almost the same.

"To me that is a fascinating response," said Cohen. It runs

"counter to everything I was expecting."

The images his machine creates appear to draw a direct connection between the inner life of the mind and the physical mechanisms of the brain.

"In a sense, this kind of functional imager . . . allows me to study covert mental activity--in essence, thought," he said.

Other recent imaging studies of brain function reinforce the relationship between conscious mental states and the physical structures of the brain:

- * When mental patients hallucinate and hear imaginary voices, those portions of their brain responsible for hearing respond as if the voices were real.

- * When people are asked to picture a map of the United States in their mind, their brains respond as if they were looking at an actual map, activating that part of the cortex responsible for vision.

- * When people are asked to picture an object and then rotate it mentally, their brains act as if the object were turning in front of them.

One surprising thing about these images of perception is that they do not show where or how the mind's eye is focused in the brain.

Subjectively, perceptions always appear seamlessly, presented in the mind as an integrated whole. But these scans showed no sign of an active anatomical area where sensory stimuli were coordinated and collated.

So how is consciousness organized in the physical structures of the brain?

To Cohen, the images suggest that consciousness is itself the act of attention.

As the brain evolved, it must have developed a way to focus itself selectively; otherwise, the constant burble of brain activity and sensory perceptions would overwhelm it. In this theory, consciousness may arise from the brain's need to concentrate, momentarily highlighting some neural activities at the expense of others.

Consciousness therefore may not be continuous, Llinas of New York University suggested.

It may flicker on and off as needed, like a spotlight flashing in the darkened theater of the mind.

An Elusive Quarry

Some scientists contend that consciousness emerges from the union of all the brain's physical properties, the way a rainbow arises from the interplay of light, suspended water droplets and air.

If that is the case, they argue, scientists will never find the human mind no matter how hard they scrutinize the brain's physical structure, any more than someone will find a rainbow in any one of its scattered parts.

Indeed, Chalmers believes scientists will discover eventually that human consciousness is an irreducible quality of the universe, like space, mass or time.

"Instead of trying to explain consciousness purely in terms of its physical processes, you should take it as a fundamental entity in its own right," he said.

"We would like a unified theory of consciousness in the same way that physicists are searching for a unified theory of matter," he said.

*

Some worry that by reducing human consciousness to its biological components, scientists will tarnish the worth of the human spirit or undercut ideas of personal responsibility and free will. But other researchers believe that a more detailed understanding of the brain will only enhance individual self-respect by affirming the uniqueness of each human mind.

"You will understand," said Nobel laureate Gerald Edelman at the Neurosciences Institute in La Jolla, "why your individuality is important."

*

In some ways, the search for the roots of human consciousness stands much of traditional scholarly inquiry on its head.

Scientists who, by temperament and training, devote their lives to the collection of objective facts about the universe around them are forced to grapple with the imponderables of the world within.

Researchers who are trained to exclude the human element from experiments discover that the human element is the experiment.

Philosophers, more accustomed to purely metaphysical speculation, are now expected to buttress their ideas with hard data.

All of them feel caught up in a uniquely human endeavor that--for better or worse--promises to alter forever humanity's sense of itself.

* * *

THE BRAIN: A WORK IN PROGRESS

A growing number of scientists are trying to study the one thing that many believe cannot be labeled, scrutinized or defined--human consciousness.

THE SEARCH FOR UNDERSTANDING

"There is something very mysterious about consciousness. Why

can objective physical systems have subjective states? It is baffling. It gets at the central idea of the soul.'

--Christoff Koch, a theoretical neurobiologist at Caltech.

THE CHALLENGE

'We are trying to understand who we are by studying the organ that allows you to understand who you are.'

--Antonio Damasio, an expert at the University of Iowa on the brain, cognition and behavior

GLOSSARY

Cortex:

Where most high-level functions associated with the mind are implemented. Some of its regions are highly specialized. For example, the occipital lobes located near the rear of the brain are associated with the visual system. The motor cortex helps coordinate all voluntary muscle movements.

Cerebral hemispheres:

Symmetrical halves of the brain. There are two occipital lobes, two parietal lobes and two frontal lobes. The two hemisphere are in continual communication with each other. Each acts as an independent parallel processor with complementary functions.

Left cerebral hemisphere:

Appears most closely associated with a conscious self. The left hemisphere, which usually manages the right side of the body, controls language and general cognitive functions. It dominates in deciding what response to make.

Right cerebral hemisphere:

Controls the left half of the body. In most people it manages nonverbal processes, such as attention, pattern recognition, line orientation and the detection of complex auditory tones.

Frontal lobes:

Located behind the forehead. They are most closely linked with making decisions and judgments.

Limbic system:

A number of interconnected brain structures linked to hormones, drives, temperature control, emotion, and, in one part, to memory formation. Neurons affecting heart rate and respiration appear concentrated in the hypothalamus and direct most of the physiological changes that accompany strong emotion.

Hippocampus:

Plays a crucial role in processing various forms of information as part of long-term memory. Damage to the hippocampus will produce global retrograde amnesia, or the inability to lay down news stores of information.

* * *

About This Series

Who are we? Where did we come from? While many scientists search for clues to these ultimate questions by probing the far reaches of the universe, others think the answers lie inside our own heads. Their probes are uncovering galaxies of neural cells, each twinkling with the brain's life forces. As it orchestrates human behavior, this symphony of electrochemical communication may indeed constitute our very essence.

Sunday: The explosion of knowledge in the field of brain development, where researchers are finding that those first few years of life are far more critical than anyone had guessed.

Monday: New technology is uncovering the brain's prominent role in emotions.

Tuesday: How brain researchers are overturning traditional ideas about mental illness.

Today: Human consciousness.

Special Internet Site

Beginning today, a special section of The Times' World Wide Webb will be devoted to information about the human brain. The full text, photos and graphics from this four-part series will be available, plus a wide range of additional information and graphics prepared especially for the Internet. Point your Web browser to:
<http://www.latimes.com/thebrain>

Copyright Los Angeles Times



NEWS • ENTERTAINMENT • DESTINATION L.A. • CLASSIFIEDS • COMMUNITIES • MARKETSPACE

 HUNTER • SPEAK OUT • HELP • CONTENTS • FIND • SO.CAL. EXCITE •  ARCHIVES • HOME

[\[NEWS\]](#) [\[ENTERTAINMENT\]](#) [\[DESTINATION L.A.\]](#) [\[CLASSIFIEDS\]](#)

[\[COMMUNITIES\]](#) [\[MARKETSPACE\]](#) [\[HUNTER\]](#) [\[SPEAK OUT\]](#) [\[HELP\]](#)

[\[CONTENTS\]](#) [\[FIND\]](#) [\[SO. CAL. EXCITE\]](#) [\[ARCHIVES\]](#) [\[HOME\]](#)

The New York Times

[Home](#)[Sections](#)[Contents](#)[Search](#)[Forums](#)[Help](#)

April 29, 1997

Consciousness Studies: From Stream to Flood



By JAMES GORMAN

NEW YORK -- At a small restaurant on West 70th Street in Manhattan, Dr. Piet Hut, an astrophysicist at the Institute for Advanced Studies in Princeton, N.J., has pushed aside his French toast to draw a quick sketch of objective and subjective reality. His drawing, on a yellow pad, shows a stick figure, a chair, the stick figure's experience of the chair, the stick figure's experience of itself and the stick figure's understanding of its experience of the chair and itself.

This is not an unusual moment in the current state of the scientific study of consciousness. Everybody from neuroscientists to philosophers to physicists is in the game. And the arena of investigation ranges from high-tech brain imaging centers to breakfast brainstorming, from monkey experiments to thought experiments.

The debate is still open as to whether the problem of consciousness -- how felt experience is connected to the physical brain -- is purely philosophical, an illusory difficulty that will be swept away by progress in neuroscience, or one that requires a revamping of all of science, starting with physics, which is what Hut is suggesting.

Whatever the approach, the stream of consciousness studies, rising for more than a decade, seems to be reaching flood stage. Dr. Daniel Dennett, a philosopher and director of the Center for Cognitive Studies at Tufts University, was asked by The Quarterly Review of Biology to review some recent books on the subject. He counted 34 in the last six years, "not including the awful ones, or those by journalists" (or those by Dennett himself), and decided there were just too many. He put the project off.

Two scientific journals on consciousness have been started in recent years. A 1995 Massachusetts Institute of Technology Press book, a 1,447-page compilation of articles titled "The Cognitive Neurosciences," is in its fourth printing and has sold 9,000 copies. The book weighs eight pounds and costs \$95. This spring and summer there are major conferences on consciousness in Montreal and California. "The scientific race for consciousness is on," writes Bernard Baars, a psychologist at the Wright Institute in Berkeley, Calif., in one of those new books, "The Theater of Consciousness" (Oxford University Press). If it is a race, it is an unconventional one in which neither the starting line nor the finish have been agreed on. The arguments are not only over answers to questions, they are over what the questions should be, and what the terms of the discussion mean. Neuroscientists, psychologists and philosophers, along with the odd physicists and anesthesiologist all have opinions and a stake in finding out what exactly goes on in that elusive, internal space where a person thinks, feels and wonders about everything from what kind of pizza to order for dinner to what exactly it means to "wonder."

The result is a field, or rather an effort to create a field, which is "boiling" with activity, says Dennett, the author of several books on consciousness, including the recent "Kinds of Minds" (Basic Books) and one of the main figures in debates about the mind. The intensity of current interest is "almost ridiculous," he said. And the disagreements are so large that the pursuers of consciousness sometimes dismiss each other with a wave of the hand.

David Chalmers, a philosopher at the University of California at Santa Cruz, and author of another new book, "The Conscious Mind: In Search of a Fundamental Theory" (Oxford University Press), says the situation reminds him of "physics 500 years before Newton" at a point when the intellectual vocabulary was not available to talk about things like gravity. Of course, even on that point there is disagreement. Baars says the situation is now like that right at the time of Newton, when gravity was beginning to be isolated and understood.

The players do agree that they disagree. Dr. John Searle, a philosopher at the

University of California at Berkeley, who panned Chalmers's new book recently in *The New York Review of Books*, a traditional jousting ground for academics, said that in the attempt to study how the physical brain and conscious mind are related, "there's no unifying principle," adding, "We're just thrashing around in the dark."

Dr. Michael Gazzaniga, a neuroscientist who started Dartmouth College's Program in Cognitive Neuroscience, said the most outrageous ideas were now reasonably discussed (he mentioned some offered by physicists), and suggested, "The reason they can get away with it is that nobody has a clue."

And among those physicists who have decided to join the fray, both Dr. Roger Penrose at Oxford and Hut say that all of science needs to be shaken up. Of course, they differ slightly in how it should be shaken up, but they both agree that modern science, as constituted, is not prepared to grapple with the conscious mind.

There are two levels of problems being solved in the pursuit of consciousness. On one level, where there is much less sound and fury, the wiring of the brain is being studied, the activity of the brain is being glimpsed in new kinds of imaging techniques and experiments are being done on animals to separate conscious from unconscious activity in the brain.

Scans of the brain, for instance, enable scientists to watch people talk to themselves, silently. Both the center for control of vocal activity and the center for analyzing speech are active while the subject says, silently, something like: "Why am I getting my brain scanned? I could be getting my hair done. What if they don't see anything happening?"

No one doubts that over the next 10 or 20 years, a much clearer picture will emerge of what happens in the brain when people, and presumably animals, have conscious experiences. But that still leaves the old, big questions that have troubled philosophers and undergraduates since there have been philosophers and undergraduates, questions like: When you talk to yourself, who talks and who listens?

And on this level, whether there is a difference between mental experience and brain activity, for instance, disagreements are dismayingly large, and on occasion blunt beyond the normal bounds of scientific discourse. In an interview, Searle said that Dennett and Chalmers were both making the "same dumb mistake" in assuming that the terms of this question were valid.

(They come up with different answers; Chalmers finds a distinction and Dennett does not.) And although his physics is respected and his books attract popular attention, almost everyone else who has a stake in consciousness dismisses Penrose out of hand. Steven Pinker at MIT, who has a much anticipated book coming out this fall from W.W. Norton, "How the Mind Works," says, "Penrose is not even in the ball park."

Penrose, on the other hand, can dismiss most of what passes for activity in neuroscience as hopeless. His contention is that something other than the computer model must be used. And he says, "The trouble with John Searle is, I don't think, although I've talked to him quite a bit, I never felt that he really grasped what a noncomputational system is. I don't think he's really got that on board yet."

Sometimes it seems like a free-for-all. Dennett said in an interview, chuckling, that "one of the really refreshing discoveries of the last few years was the advent of physics getting into the act," because it proved that "there's a discipline that is even more ignorant about the brain and even more arrogant than philosophy."

Among the several positions held about consciousness and the self, Dennett and many others believe consciousness is simply what the brain does, a process, like life or growth, not a great philosophical problem. Gazzaniga believes that it is indeed something the brain does, but he offers a location for the self, an "interpreter module" in the left cerebral cortex. Some philosophers say the whole notion that activity in the brain can account for subjective experience is naive and unacceptable.

The headier discussions are worrisome to some more traditional neuroscientists. Dr. Francis Crick, who won the Nobel Prize for his discovery with Dr. James Watson of the structure of DNA, has attracted many neuroscientists to the field with his forays into the subject of the mind. But he said at a recent talk that he wished journalists would stop writing about consciousness.

Dr. Christof Koch, a neuroscientist at the California Institute of Technology, said the interest in consciousness "might be too much of a good thing." He and others say that there is essentially no progress on the more difficult philosophical problems of self-awareness, and that the concentration should be in hard experimental studies on subjects like visual perception.

"We're going to be called to the table in 5 or 10 years," said Koch, and the

current fascination with the mind could well fade, as it has in the past, and research support with it. Chalmers said, "If the results aren't delivered in a few years someone might say, well, there was this consciousness craze in the 1990's and look what came of it."

If all goes well, said Chalmers, himself a philosopher, scientific interest in consciousness will continue to increase, and the problem will move further and further away from philosophy as it gets closer to being solved.

Chalmers remarked, "I'd hate for it to go the other way."

Copyright 1997 The New York Times Company

Online papers on consciousness, part 1: Philosophy of consciousness

Compiled by [David Chalmers](#)

Part 1 of [Online papers on consciousness](#)

- Part 1: Philosophy of consciousness
- [Part 2: Other philosophy of mind](#)
- [Part 3: Science of Consciousness](#)

TABLE OF CONTENTS [368 papers]

- [The concept of consciousness](#) [26 papers]
- [The explanatory gap](#) [33]
- [Materialism and dualism](#) [29]
- [The knowledge argument](#) [16]
- [Materialism and modality](#) [29]
- [Metaphysics of consciousness](#) [33]
- [Panpsychism](#) [11]
- [Zombies](#) [18]
- [Qualia, color, etc.](#) [39]
- [The contents of consciousness](#) [20]
- [Representationalism](#) [17]
- [Consciousness as higher-order thought](#) [23]
- [Introspection, self-knowledge, and self-consciousness](#) [28]
- [The unity of consciousness](#) [11]
- [The function of consciousness](#) [18]
- [Philosophy of consciousness \(misc.\)](#) [23]

The Concept of Consciousness

- Michael Antony, [Outline of a general methodology for consciousness research](#)
- Michael Antony, [Is "consciousness" ambiguous?](#)
- Ned Block, [On a confusion about a function of consciousness](#)
- David Chalmers, [Availability: The cognitive basis of experience?](#)
- Austen Clark, [Phenomenal consciousness so-called](#)
- Ronnie de Sousa, [Twelve varieties of subjectivity](#)
- Eric Gillett, [Searle and the "deep unconscious"](#) (and Dan Lloyd's [commentary](#))
- Alvin Goldman, [Consciousness, folk psychology, and cognitive science](#)

- Ted Honderich, [Consciousness as existence](#)
- Ted Honderich, [Consciousness as existence again](#)
- William James, [Does 'consciousness' exist?](#)
- Douglas Katz, [Minimally conscious states](#)
- Eric Lormand, [Consciousness](#)
- Eric Lormand, [Nonphenomenal consciousness](#)
- William Lycan, [The plurality of consciousness](#)
- Gregory Mulhauser, [What is self-awareness?](#)
- David Rosenthal, [Consciousness and sensation](#)
- David Rosenthal, [How many kinds of consciousness?](#)
- David Rosenthal, [The kinds of consciousness](#)
- David Rosenthal, [State consciousness and transitive consciousness](#)
- Aaron Sloman, [Notes on consciousness](#)
- Aaron Sloman, [What is it like to be a rock?](#)
- Nigel Thomas, [Imagination, eliminativism, and the pre-history of consciousness](#)
- Max Velmans, [Defining consciousness](#)

The Explanatory Gap

- Marcus Arvan, [Out with qualia and in with consciousness: Why the hard problem is a myth](#)
- Giorgio Ascoli, [Is it already time to give up on a science of consciousness?](#)
- Ned Block, [The harder problem of consciousness](#)
- Ned Block and Robert Stalnaker, [Conceptual analysis and the explanatory gap](#)
- Peter Carruthers, [Consciousness: Explaining the phenomena](#)
- David Chalmers, [Facing up to the problem of consciousness](#)
- David Chalmers, [Moving forward on the problem of consciousness](#)
- David Chalmers, [The puzzle of conscious experience](#)
- David Chalmers and Frank Jackson, [Conceptual analysis and reductive explanation](#)
- Patricia Churchland, [The hornswoggle problem](#) (and Chalmers' [reply](#))
- Austen Clark, [How to respond to philosophers on raw feels](#)
- Austen Clark, [I am Joe's explanatory gap](#)
- Tom Clark, [Function and phenomenology: Closing the explanatory gap](#)
- David de Leon, [The limits of thought and the mind-body problem](#)
- Daniel Dennett, [Facing backwards on the problem of consciousness](#) (and Chalmers' [reply](#))
- James Garvey, [What does McGinn think we cannot know?](#)
- Brie Gertler, [The explanatory gap is not an illusion](#)
- Brie Gertler, [Explanatory reduction, conceptual analysis, and conceivability arguments about the mind](#)

- Marion Gothier, [Consciousness, science, and the nature of explanation](#)
- David Hilbert & Mark Kalderon, [Color and the inverted spectrum](#)
- James Hopkins, [Mind as metaphor: A physicalistic approach to the problem of consciousness](#)
- Steven Horst, [Evolutionary explanation and the hard problem of consciousness](#)
- Marc Krellenstein, [Unsolvability problems, visual imagery, and explanatory satisfaction](#)
- Eric Lormand, [The explanatory stopgap](#) (and [appendix](#))
- Bruce MacLennan, [The elements of consciousness and their neurodynamical correlates](#)
- Ron McClamrock, [Irreducibility and subjectivity](#)
- Barbara Montero, [The epistemic/ontic divide](#)
- David Papineau, [Mind the gap](#)
- Teed Rockwell, [The hard problem is dead: Long live the hard problem](#)
- Michael Tye, [Phenomenal consciousness: The explanatory gap as cognitive illusion](#)
- Max Velmans, [The relation of consciousness to the material world](#)

Materialism and Dualism

- Torin Alter, [Nagel on imagination and physicalism](#)
- John Beloff, [What are minds for?](#)
- John Beloff, [The mind-brain problem](#)
- Glenn Braddock, [Against Chalmers' epiphenomenalism](#)
- Selmer Bringsjord, [Searle on the brink](#)
- David Chalmers, [Consciousness and its place in nature](#)
- Andrew Chrucky, [Critique of Wilfrid Sellars' materialism](#)
- Rene Descartes, [Meditations on First Philosophy](#)
- John Hubbard, [Parsimony and the mind](#)
- Peter J. King, [Dualism: An empirical test](#)
- E.J. Lowe, [Self, agency, and mental causation](#)
- Peter Lloyd, [Is the mind physical? Dissecting conscious brain tissue](#)
- Marvin Minsky, [Minds are simply what brains do](#)
- Marvin Minsky, [Matter, minds, and models](#)
- Barbara Montero, [The body problem](#)
- Barbara Montero, [Post-physicalism](#)
- Hans Moravec, [Dualism through reductionism](#)
- David Papineau, [Thinking about Consciousness](#)
- David Papineau, [The rise of physicalism](#)
- Thomas Polger, [Escaping the epiphenomenal trap](#)
- Karl Popper, [Language and the body-mind problem: A restatement of interactionism](#)
- Karl Popper, [A note on the body-mind problem](#)
- Howard Robinson, [Materialism in the philosophy of mind](#)

- William Robinson, [Epiphenomenalism](#)
- David Rosenthal, [The identity theory](#)
- A.J. Rudd, [What it's like and what's really wrong with physicalism: A Wittgensteinian perspective](#)
- Wilfrid Sellars, [A note on Popper's argument for dualism](#)
- Wilfrid Sellars, [Is consciousness physical?](#)
- Daniel Stoljar, [Two conceptions of the physical](#)

The Knowledge Argument

- Torin Alter, [The knowledge argument](#)
- Torin Alter, [Know-how, ability, and the ability hypothesis](#)
- Laurence Bonjour, [What is it like to be human \(instead of a bat\)](#)
- Alex Byrne, [Something about Mary](#)
- Max Deutsch, [Subjective physical facts](#)
- Brie Gertler, [A defense of the knowledge argument](#)
- Frank Jackson, [Epiphenomenal qualia](#)
- William Lycan, [Perspectival representation and the knowledge argument](#)
- Thomas Nagel, [What is it like to be a bat?](#)
- David Rosenthal, [Reductionism and knowledge](#)
- Fredrik Stjernberg, [Not so epiphenomenal qualia](#)
- Nigel Thomas, [Mary doesn't know science: On misconceiving a science of consciousness](#)
- Michael Tye, [Knowing what it is like: The ability hypothesis and the knowledge argument](#)
- Adam Vinuesa, [The knowledge argument](#) (and [another](#))
- Sven Walter, [Terry, Terry, quite contrary](#)

Materialism and Modality

- Murat Aydede & Guven Guzeldere, [Consciousness, conceivability arguments, and perspectivalism](#)
- Katalin Balog, [Conceivability arguments, or revenge of the zombies](#)
- Alex Byrne, [Cosmic hermeneutics](#)
- David Chalmers, [Insentience, indexicality, and intensions \(review of Perry\)](#)
- David Chalmers, [Materialism and the metaphysics of modality](#)
- David Chalmers, [Mind and modality](#)
- David Chalmers, [The nature of epistemic space](#)
- David Chalmers, [Does conceivability entail possibility?](#)
- Adam Elga, [A conceivability argument](#)
- Ivan Havel, [Living in conceivable worlds](#)
- Christopher Hill, [Chalmers on the apriority of modal knowledge](#)

- Mark Johnston, [It necessarily ain't so](#)
- Navin Kartik, [In the hands of zombies](#)
- Brian Loar, [Phenomenal states](#)
- Farid Masrour, [Are zombies positively conceivable?](#)
- Farid Masrour, [Taking Kripke seriously](#)
- Wallace Matson, [Logical possibility, laws of nature, and mind in the history of philosophy](#)
- Thomas Nagel, [Conceiving the impossible and the mind-body problem](#)
- Matthew Phillips, [Why positive and negative conceivability can't save the conceivability-possibility link](#)
- Paul Raymore, [A materialist response to David Chalmers' *The Conscious Mind*](#)
- Robert Stalnaker, [On considering a possible world as actual](#)
- Ralph Wedgwood, [The price of nonreductive physicalism](#)
- Stephen White, [Why the property dualism argument won't go away](#)
- Crispin Wright, [The conceivability of naturalism](#)
- Stephen Yablo, [Concepts and consciousness](#)
- Stephen Yablo, [Coulda, woulda, shoulda](#)
- Stephen Yablo, [Modal rationalism and logical empiricism: Some similarities](#)
- Stephen Yablo, [Textbook Kripkeanism and the open texture of language](#)
- Stephen Yablo, [Grokking pain](#)

Metaphysics of Consciousness

- Sophie Allen, [A space oddity: McGinn on consciousness and space](#)
- John Bolender, [Factual phenomenalism: A supervenience theory](#)
- C.D. Broad, [Mind and Its Place in Nature](#)
- Dan Bruiger, [The rise and fall of reality](#)
- David Chalmers, [Consciousness and cognition](#)
- Fergus Duniho, [The mind/body problem and its solution](#)
- Ed Feser, [Hayek's solution to the mind-body problem](#)
- Richard Gale, [William James on the misery and glory of consciousness](#)
- Rom Harre, [Nagel's challenge and the mind-body problem](#)
- Dan Hutto, [An ideal solution to the problem of consciousness](#)
- Ted Honderich, [Consciousness, neural functionalism, and real subjectivity](#)
- Ted Honderich, [Functionalism, identity theories, the union theory](#)
- Piet Hut & Bas van Fraassen, [Elements of reality: A dialogue](#)
- William James, [A world of pure experience](#)
- Mark Johnston, [The Manifest, Chapter 1](#) (also [Chapter 5](#), [Chapter 7](#))
- Peter Lloyd, [Berkelian ontology as a fundamental approach to consciousness](#)

- Michael Lockwood, [The grain problem](#)
- Ron McClamrock, [Irreducibility and subjectivity](#)
- Colin McGinn, [Consciousness and space](#)
- Paul Meehl & Wilfrid Sellars, [The concept of emergence](#)
- Anil Mitra, [Problems in the science and philosophy of mind and consciousness](#)
- David Newman, [Chaos, emergence, and the mind-body problem](#)
- Stephen Pepper, [Emergence](#)
- John Perry, [Knowledge, Possibility, and Consciousness](#)
- Gregg Rosenberg, [A Place for Consciousness: Probing the Deep Structure of the Natural World](#)
- Wilfrid Sellars, [Empiricism and the philosophy of mind](#)
- Galen Strawson, [Realistic materialism](#)
- Max Velmans, [A natural account of phenomenal consciousness](#)
- Max Velmans, [Consciousness, brain, and the physical world](#)
- Max Velmans, [Goodbye to reductionism](#)
- Adam Vinuesa, [Functionalism with and without circularity](#)

Panpsychism

- Charles Birch, [Why I became a panexperientialist](#)
- Peter Farleigh, [Whitehead's even more dangerous idea](#)
- Liane Gabora, [Amplifying phenomenal information: Toward a fundamental theory of consciousness](#)
- Piet Hut and Roger Shepard, [Turning the "hard problem" upside down and sideways](#)
- David Pearce, [Cosmic consciousness for tough minds](#)
- David Pearce, [Naturalistic panpsychism](#)
- Gregg Rosenberg, [Rethinking nature: On the possibility of a benign panpsychism](#)
- William Seager, [Consciousness, information, and panpsychism](#)
- William Seager, [Panpsychism](#)
- T.L.S. Sprigge, [Panpsychism](#)
- Peter Unger, [The mystery of the physical and the matter of qualities](#)

Zombies

- Selmer Bringsjord, [In defense of impenetrable zombies](#)
- Selmer Bringsjord, [The zombie attack on the computational conception of mind](#)
- David Chalmers, [Self-ascription without qualia: A case-study](#)
- Allin Cottrell, [Sniffing the camembert: On the conceivability of zombies](#)
- Daniel Dennett, [The unimagined preposterousness of zombies](#)
- Daniel Dennett, [The zombic hunch: Extinction of an intuition?](#)

- Owen Flanagan & Tom Polger, [Zombies and the function of consciousness](#)
- Stevan Harnad, [Why and how we are not zombies](#)
- Larry Hauser, [Revenge of the zombies](#)
- Jaron Lanier, [You can't argue with a zombie](#)
- Peter Marton, [Zombies versus materialists: The battle for conceivability](#)
- Chris Mathieson, [Reining\ in Chalmers: On the logical possibility of zombies](#)
- John McCarthy, [Todd Moody's zombies](#)
- Todd Moody, [Conversations with zombies](#)
- Gregory Mulhauser, [Zombies](#)
- Tom Polger, [Zombies](#)
- Nigel Shardlow, [Zombies](#)
- Nigel Thomas, [Zombie killer](#)

Qualia, Color, etc.

- Torin Alter, [Qualia](#)
- Murat Aydede, [Naturalism, introspection, and direct realism about pain](#)
- Murat Aydede, [An analysis of pleasure vis-a-vis pain](#)
- Murat Aydede & Guven Guzeldere, [Some foundational problems in the scientific study of pain](#)
- Jose Luis Bermudez, [Categorizing qualitative states: Some problems](#)
- Alex Byrne, [Do colours look like dispositions?](#)
- Alex Byrne & David Hilbert, [Colors and reflectances](#)
- Alex Byrne & David Hilbert, [Color realism and color science](#)
- David Chalmers, [Absent qualia, fading qualia, dancing qualia](#)
- Andrew Chrucky, [The alleged fallacy of the sense-datum inference](#)
- Austen Clark, [A physicalist theory of qualia](#)
- Austen Clark, [A subjectivist reply to spectrum inversion](#)
- Austen Clark, [Qualia and the psychophysical explanation of color perception](#)
- Austen Clark, [Spectrum inversion and the color solid](#)
- Austen Clark, [The particulate instantiation of homogeneous pink](#)
- David Cole, [Inverted spectrum arguments](#)
- David de Leon, [The qualities of qualia](#)
- Daniel Dennett, [Instead of qualia](#)
- Daniel Dennett, [Lovely and suspect qualities](#)
- Daniel Dennett, [Quining qualia](#)
- Richard Gregory, [Peculiar qualia](#)
- Richard Gregory, [What do qualia do?](#)
- David Hilbert, [What is color vision?](#)
- Nicholas Humphrey, [The privatization of sensation](#)

- Eric Lormand, [Qualia! \(Now showing at a theater near you\)](#)
- Zoltan Jakab, [Metameric surfaces: The ultimate case against color physicalism and representational theories of phenomenal consciousness](#)
- Stephen Palmer, [Color, consciousness, and the isomorphism constraint](#)
- Eugene Park, [Against Dennett's eliminativism: Preserving qualia as a coherent concept](#)
- Tony Pitson, [The dispositional account of colour](#)
- Teed Rockwell, [Experience and sensation: Sellars and Dewey on the non-cognitive aspects of mental life](#)
- David Rosenthal, [The colors and shapes of visual experiences](#)
- David Rosenthal, [Color, mental location, and the visual field](#)
- Mark Sharlow, [Visual cortical feedback and the ineffability of colors](#)
- Sydney Shoemaker, [Color, subjective reactions, and qualia](#)
- Sydney Shoemaker, [The phenomenal character of experience](#)
- Wolfgang Spohn, [The character of color terms: A materialist view](#)
- Robert Stalnaker, [Comparing qualia across persons](#)
- Michael Tye, [Qualia](#)

The Contents of Consciousness

- Kent Bach, [How can experiences find their objects?](#)
- Jose Luis Bermudez, [Nonconceptual content: From perceptual experience to subpersonal computational states](#)
- Jose Luis Bermudez and Fiona McPherson, [Nonconceptual content and the nature of perceptual experience](#)
- Peter Carruthers, [Conscious thinking: Language or elimination?](#)
- Austen Clark, [Sensing and reference](#)
- Daniel Dennett, [Seeing is believing--or is it?](#)
- Shaun Gallagher, [The Molyneux Problem](#)
- Ted Honderich, [Seeing things](#)
- Tomis Kapitan, [Vision, vector, veracity](#)
- Sean Kelly, [On what we see: The phenomenology and analysis of perceptual experience](#)
- Sean Kelly, [Demonstrative concepts and experience](#)
- Sean Kelly, [The non-conceptual content of perceptual experience: situation-dependence and fineness of grain](#)
- Sean Kelly, [What makes perceptual content nonconceptual?](#)
- Uriah Kriegel, [The intentionality of conscious experience and mind-relative content](#)
- Brian Loar, [Phenomenal intentionality as the basis of mental content](#)
- Michael Martin, [The transparency of experience](#)

- Fiona McPherson, [Perfect pitch and the content of experience](#)
- Mark Rowlands, [Two dogmas of consciousness](#)
- Norman Teng, [The depictive nature of visual mental imagery](#)
- Nigel Thomas, [Coding dualism: Conscious thought without Cartesianism or computationalism](#)
- Wayne Wright, [McDowell, demonstrative concepts, and nonconceptual content](#)

Representationalism

- Kent Bach, [Engineering the mind](#)
- Ned Block, [Mental paint](#)
- Ned Block, [Sexism, ageism, racism, and the nature of consciousness](#)
- Alex Byrne, [Intentionalism defended](#)
- David Cole, [Dretske on naturalizing the mind](#)
- Guven Guzeldere & Murat Aydede, [On the relation between phenomenal and representational properties](#)
- Benj Hellie, [Presence to the Mind](#)
- Robert Kirk, [Why ultra-externalism goes too far](#)
- Brendar Lalor, [Intentionality and qualia](#)
- William Lycan, [Representational theories of consciousness](#)
- William Lycan, [The case for phenomenal externalism](#)
- Daniel Stoljar, [What what it's like isn't like](#)
- Michael Tye, [What what it's like is really like](#)
- Michael Tye, [Inverted Earth, Swampman, and representationism](#)
- Adam Vinuesa, [Sensations and the language of thought](#)
- Wayne Wright, [Tye, tree-rings, and representation](#)

Consciousness as Higher-order Thought

- Derek Browne, [Consciousness is not HOT](#)
- Alex Byrne, [Some like it HOT: Consciousness and higher-order thought](#)
- Peter Carruthers, [Higher-order theories of consciousness](#)
- Peter Carruthers, [Natural theories of consciousness](#)
- Peter Carruthers, [Precis of *Phenomenal Consciousness: A Naturalistic Theory*](#) (and [commentaries](#); also [here](#))
- Peter Carruthers, [Recognitional concepts and phenomenal properties](#)
- David Cole, [Sense and sentience](#)
- Eric Lormand, [Inner sense until proven guilty](#)
- William Lycan, [A simple argument for a higher-order representation theory of consciousness](#)
- William Lycan, [Consciousness as internal monitoring](#)

- David Rosenthal, [A theory of consciousness](#)
- David Rosenthal, [Apperception, sensation, and dissociability](#)
- David Rosenthal, [Consciousness, content, and metacognitive judgments](#) (and [response to commentators](#))
- David Rosenthal, [Consciousness and metacognition](#)
- David Rosenthal, [Content, interpretation, and consciousness](#)
- David Rosenthal, [Higher-order thoughts and the appendage theory of consciousness](#)
- David Rosenthal, [Perceptual and cognitive models of consciousness](#)
- David Rosenthal, [State consciousness and what it's like](#)
- William Seager, [On dispositional HOT theories of consciousness](#)
- Adam Vinuesa, [Inner monitoring and phenomenal consciousness](#)

Introspection, Self-Knowledge, and Self-Consciousness

- Murat Aydede & Güven Güzeldere, [Information, introspection, and phenomenal consciousness](#)
- Murat Aydede, [Is introspection inferential?](#)
- Tim Bayne, [Chalmers on the justification of phenomenal judgments](#)
- Jose Luis Bermudez, [Nonconceptual self-awareness and the paradox of self-consciousness](#)
- Jose Luis Bermudez, [The Paradox of Self-Consciousness \(precis\)](#) (and [discussion](#), and [more discussion](#))
- Andrew Brook, [Externalism and the varieties of self-awareness](#)
- Andrew Brook, [Kant, self-awareness, and self-reference](#)
- David Chalmers, [The content and epistemology of phenomenal belief](#)
- Fred Dretske, [The mind's awareness of itself](#)
- Shaun Gallagher & Andrew Meltzoff, [The earliest sense of self: Merleau-Ponty and recent developmental studies](#)
- Shaun Gallagher, [Non-perceptual self-awareness](#)
- Shaun Gallagher, [Self-reference and schizophrenia: A cognitive model of immunity to error through misidentification](#)
- Brie Gertler, [Introspecting phenomenal states](#)
- Brie Gertler, [The mechanics of self-knowledge](#)
- George Graham, [Self-consciousness, psychopathology, and realism about self](#)
- Susan Hurley, [Nonconceptual self-consciousness and agency: Perspective and access](#)
- William James, [The consciousness of self](#)
- Tomis Kapitan, [The ubiquity of self-awareness](#)
- Eric Lormand, [Phenomenal illusions](#)
- Christopher Peacocke, [Conscious attitudes, attention, and self-knowledge](#)
- John Perry, [Myself and "I"](#)
- Jay Rosenberg, [Perception vs. inner sense: A problem about direct awareness](#)

- David Rosenthal, [Introspection](#)
- Eric Schwitzgebel, [How well do we know our own conscious experience? The case of imagery](#)
- Eric Schwitzgebel, [Why did we think we dreamed in black and white?](#)
- Eric Schwitzgebel & Michael Gordon, [How well do we know our own conscious experience? The case of human echolocation](#)
- Robert Van Gulick, [Inward and upward: Reflection, introspection, and self-awareness](#)

The Unity of Consciousness

- Tim Bayne, [The unity of consciousness: Clarification and defence](#)
- Tim Bayne, [Review of Dainton's *Stream of Consciousness*](#)
- Tim Bayne and David Chalmers, [What is the unity of consciousness?](#)
- Andrew Brook, [The unity of consciousness](#)
- Andrew Brook, [Judgments and drafts eight years later](#)
- Rodney Cotterill, [On the unity of conscious experience](#)
- Nicholas Humphrey, [One-self: A meditation on the unity of consciousness"](#)
- Gerard O'Brien & Jon Opie, [The disunity of consciousness](#)
- Gerard O'Brien & Jon Opie, [The multiplicity of consciousness and the emergence of the self](#)
- Gregg Rosenberg, [The boundary problem for phenomenal individuals](#)
- Christoph von der Malsburg, [The coherence definition of consciousness](#)

The Function of Consciousness

- Bernard Baars, [The functions of consciousness](#)
- J.M. Baldwin, [Consciousness and evolution](#) (and [sequel](#))
- Ned Block, [On a confusion about a function of consciousness](#)
- Selmer Brinsgjord & Ron Noel, [Why did evolution engineer consciousness?](#)
- Peter Carruthers, [The evolution of consciousness](#)
- Fred Dretske, [What good is consciousness?](#)
- Owen Flanagan & Thomas Polger, [Zombies and the function of consciousness](#)
- David Hilbert, [Why have experiences?](#)
- Nicholas Humphrey, [The uses of consciousness](#)
- George Mandler, [Consciousness redux](#)
- Thomas Polger & Owen Flanagan, [Explaining the evolution of consciousness: The other hard problem](#)
- Aaron Sloman, [The evolution of what?](#)
- Arnold Tannenbaum, [The sense of consciousness](#)
- Max Velmans, [Is human information processing conscious?](#) (plus [first](#), [second](#), and [third](#) reply to commentators)

Philosophy of Consciousness (Misc.)

- Renate Bartsch, [*Consciousness at the Cutting Edge between World and Brain*](#)
 - Ned Block, [Troubles with functionalism](#)
 - Ned Block, [Paradox and cross purposes and recent work on consciousness](#)
 - Selmer Bringsjord, [Explaining phi without Dennett's exotica: Good ol' computation suffices](#)
 - David Chalmers, [Consciousness and cognition](#)
 - Andy Clark, [I am John's brain](#)
 - Tom Clark, [Death, nothingness, and subjectivity](#)
 - Daniel Dennett, [Are we explaining consciousness yet?](#)
 - Daniel Dennett, [Is perception the "leading edge" of memory?](#)
 - Matthew Elton, [Consciousness: Only at the personal level](#)
 - Susan Hurley, [Precis of *Consciousness in Action*](#) (and [commentaries](#))
 - Dan Hutto, [Consciousness demystified: A Wittgensteinian critique of Dennett's project](#)
 - Julian Jaynes, [The problem of consciousness](#)
 - Eric Lormand, [Steps toward a science of consciousness?](#)
 - Gerard O'Brien & Jon Opie, [A defense of Cartesian materialism](#)
 - Thomas Nagel, [The mind wins](#)
 - David Papineau, [Theories of consciousness](#)
 - David Rosenthal, [Multiple drafts and facts of the matter](#)
 - John Searle, [The problem of consciousness](#)
 - John Searle, [Consciousness and the philosophers](#) (and Chalmers' [reply](#))
 - Charles Siewert, [Precis of *The Significance of Consciousness*](#) (and [commentaries](#))
-

Go to:

- [Online papers on consciousness](#)
- [David Chalmers' home page](#)

Online papers on consciousness, part 2: Other Philosophy of Mind

Compiled by [David Chalmers](#)

Part 2 of [Online papers on consciousness](#)

- [Part 1: Philosophy of Consciousness](#)
- Part 2: Other Philosophy of Mind
- [Part 3: Science of Consciousness](#)

TABLE OF CONTENTS [433 papers]

- [Language and thought](#) [27]
- [Propositional attitudes](#) [14]
- [Theories of content](#) [23]
- [Internalism and externalism](#) [32]
- [Concepts](#) [16]
- [Mental content \(misc.\)](#) [24]
- [Folk psychology](#) [35]
- [Metaphysics of mind](#) [46]
- [The self and personal identity](#) [37]
- [Free will](#) [31]
- [Philosophy of mind \(misc.\)](#) [12]
- [The Chinese Room](#) [16]
- [Godel's theorem and AI](#) [22]
- [The Turing test](#) [18]
- [Philosophy of connectionism](#) [26]
- [Computationalism](#) [25]
- [Philosophy of AI \(misc.\)](#) [29]
- [Consciousness and AI](#) (in part 3)

Language and Thought

- Murat Aydede, [The language of thought hypothesis](#) (see also [here](#))
- Peter Carruthers, [Thinking in language?: Evolution and a modularist possibility](#)
- Peter Carruthers, [The cognitive functions of language](#)
- David Chalmers, [Is there synonymy in Ockham's mental language](#)
- Andrew Chrucky, [Sellars on language and thought](#)
- Andy Clark, [Magic words: How language augments human computation](#)

- David Cole, [Pinker on the thinker: Against mentalese monopoly](#)
- David Cole, [Hearing yourself think: Natural language, inner speech, and thought](#)
- Daniel Dennett, [Verbal language as a communicative system](#)
- Daniel Dennett, [The role of language in intelligence](#)
- Daniel Dennett, [Making tools for thinking](#)
- Daniel Dennett, [Reflections on language and mind](#)
- Richard DeWitt, [Vagueness, semantics, and the language of thought](#)
- Jerry Fodor, [Why there still has to be a language of thought](#)
- Jerry Fodor, [Doing without *What's Within*](#) (and Fiona Cowie's [response](#); see also [these](#))
- Christopher Gauker, [Language and thought](#)
- Larry Hauser and Barbara Abbott, [Natural language and thought](#)
- Frank Jackson, [Learning from Locke on voluntary signs](#)
- Larry Kaye, [The language of thought](#)
- Ruth Millikan, [The language-thought partnership: A bird's eye view](#)
- Brent Silby, [Revealing the language of thought](#)
- Aaron Sloman, [The primacy of non-communicative language](#)
- Dan Sperber, [Understanding verbal understanding](#)
- Dan Sperber & Deirdre Wilson, [The mapping between the mental and the public lexicon](#)

Propositional Attitudes

- Joao Branquinho, [The problem of cognitive dynamics](#)
- David Braun, [Russellianism and psychological generalizations](#) (and [explanation](#) and [prediction](#))
- Curtis Brown, [Direct and indirect belief](#)
- Curtis Brown, [What is a belief state?](#)
- Austen Clark, [Beliefs and desires incorporated](#)
- Daniel Dennett, [Do animals have beliefs?](#)
- Henry Jackman, [Truth, rationality, and humanity](#)
- Joe Lau, [Possible worlds semantics for belief sentences](#)
- John Perry, [Circumstantial attitudes and benevolent cognition](#)
- John Perry, [Rip Van Winkle and other characters](#)
- Eric Schwitzgebel, [A phenomenal, dispositional account of belief](#)
- Dan Sperber, [Intuitive and reflective beliefs](#)

Theories of Content

- Fred Adams, [Fodor's asymmetrical causal dependency theory of meaning](#)
- Ned Block, [Conceptual role semantics](#)
- Alex Byrne, [Interpretivism](#)

- David Cole, [Natural language and natural meaning](#)
- Adrian Cussins, [Content, embodiment, and objectivity](#)
- Daniel Dennett, [Intentional systems](#)
- Daniel Dennett, [Evolution, error, and intentionality](#)
- Daniel Dennett, [Intentionality](#)
- Daniel Dennett, [The interpretation of texts, people, and other artifacts](#)
- Crawford Elder, [What vs. how in naturally selected representation](#)
- Chris Eliasmith, [How neurons mean: A neurocomputational theory of representational content](#)
- Hartry Field, [Deflationist views of meaning and content](#) (and [postscript](#))
- Gilbert Harman, [\(Nonsolipsistic\) conceptual role semantics](#)
- Henry Jackman, [Conventionalism, objectivity, and constitution](#)
- Brendar Lalor, [Swampman, etiology, and content](#)
- Gerard O'Brien & Jon Opie, [Notes towards a structuralist theory of mental representation](#)
- Ruth Millikan, [Theories of mental content](#)
- Ruth Millikan, [Naturalizing intentionality](#)
- Ruth Millikan, [On swampkinds](#)
- Ruth Millikan, [Reading Mother Nature's mind](#)
- Ruth Millikan, [What has natural information have to do with intentional representation?](#)
- Mark Rowlands, [Teleosemantics](#)
- Tad Zawidzki, [How to criticize Millikan: The problem of mythological content](#)

Internalism and Externalism

- Louise Antony, [Equal rights for swamppersons](#)
- Louise Antony, [What are you thinking? Character and content in the language of thought](#)
- Kent Bach, [Content: Wide vs. narrow](#)
- Paul Boghossian, [What the externalist can know a priori](#)
- Curtis Brown, [Belief states and narrow content](#)
- Jessica Brown, [The incompatibility of anti-individualism and privileged access](#)
- David Buller, [Individualism and evolutionary psychology \(or: In defense of 'narrow' functions\)](#)
- David Buller, ["Narrow"-mindedness breeds inaction](#)
- David Chalmers, [The components of content](#)
- Andy Clark & David Chalmers, [The extended mind](#)
- Bryan Frances, [Twin Earth thought experiments](#)
- Joseph Fulda, ["The extended mind"--extended](#)
- Susan Hurley, [Vehicles, contents, conceptual structure, and externalism](#)
- Henry Jackman, [Individualism and interpretation](#)
- Henry Jackman, [Deference and self-knowledge](#)

- Brendar Lalor, [It is what you think: Intentional potency & anti-individualism](#)
- Joe Lau, [Three motivations for narrow content](#)
- Joe Lau, [Externalism about mental content](#)
- Peter Ludlow, [Externalism, self-knowledge, and the prevalence of slow switching](#)
- Peter Ludlow, [Social externalism, self-knowledge, and memory](#)
- Peter Ludlow, [Social externalism and memory: A problem](#)
- Peter Ludlow, [On the relevance of slow switching](#)
- Luca Malatesti, [Externalism and the knowledge of mental states](#)
- Ron McClamrock, [Methodological individualism considered as a constitutive principle of scientific inquiry](#)
- Brian McLaughlin and Michael Tye, [Is content-externalism compatible with privileged access?](#)
- Karen Neander, [The narrow and the normative](#)
- John Perry, [Broadening the mind](#)
- Sarah Sawyer, [An externalist account of introspective knowledge](#)
- Sarah Sawyer, [Privileged access to the world](#)
- Michael Tye, [Externalism and memory](#)
- Alberto Voltolini, [Internalism and externalism](#)
- Wayne Wright, [Individualism, behavior, and Marr's theory of vision](#)

Concepts

- Murat Aydede, [Fodor on concepts and Frege puzzles](#)
- Kent Bach, [Review of *Concepts: Where Cognitive Science Went Wrong*](#)
- Jose Luis Bermudez, [Naturalism and conceptual norms](#)
- Harold Brown, [Sellars, concepts, and conceptual change](#)
- Harold Brown, [Conceptual comparison and conceptual innovation](#)
- Jerry Fodor, [Concepts: Where Cognitive Science Went Wrong](#)
- Jerry Fodor, [There are no recognitional concepts, not even RED](#) (and [part 2](#))
- Jerry Fodor & Ernie Lepore, [The red herring and the pet fish: Why concepts still can't be prototypes](#)
- Jay Garfield, [Thought as language: A metaphor too far](#)
- Ulrike Haas-Spohn & Wolfgang Spohn, [Concepts are beliefs about essences](#)
- Larry Hauser, [Doing without mentalese](#)
- Ruth Millikan, [A common structure for concepts of individuals, stuffs, and real kinds: More mama, more milk, and more mouse](#) (and [response](#) to critics)
- Ruth Millikan, [How we make our ideas clear: Empiricist epistemology for empirical concepts](#)
- Ruth Millikan, [A more plausible kind of "recognitional concept"](#)
- Stephen Schiffer, [Meanings and concepts](#)
- Paul Thagard, [Coherent and creative conceptual combinations](#)

Mental Content, (Misc.)

- Murat Aydede, [Syntax, content, and functionalism: What is wrong with the syntactic theory of mind](#)
- Jose Luis Bermudez, [Syntax, semantics, and levels of explanation](#)
- Ned Block, [Mental and semantic holism](#)
- Adrian Cussins, [Subjectivity, objectivity, and theories of reference in Evans' theory of thought](#)
- Daniel Dennett, [Things about things](#)
- Hartry Field, [Remarks on content and its role in explanation](#)
- Rick Grush, [Skill and spatial content](#)
- David Hilbert, [Content, intention, and explanation](#)
- Steven Horst, [Notions of 'representation' in philosophy and empirical research](#)
- Steven Horst, [Symbols and computation: A critique of the computational theory of mind](#)
- Henry Jackman, [James' pragmatic account of intentionality and truth](#)
- Henry Jackman, [Moderate holism and the instability thesis](#)
- Henry Jackman, [Holism, meaning, and context](#)
- Sean Kelly, [Grasping at straws: Motor intentionality and the cognitive science of skillful action](#)
- Eric Lormand, [How to be a meaning holist \(and appendix\)](#)
- Eric Lormand, [How to be a meaning atomist](#)
- Ruth Millikan, [The myth of mental indexicals](#)
- Ruth Millikan, [Pushmi-pullyu representations](#)
- John Perry, [Intentionality and its puzzles](#)
- John Perry and David Israel, [Fodor and psychological explanation](#)
- Stephen Schiffer, [A paradox of meaning](#)
- Stephen Schiffer, [An introduction to content and its role in explanation](#)
- Aaron Sloman, [Toward a general theory of representations](#)
- Robert Stalnaker, [What might nonconceptual content be?](#)

Folk Psychology

- Angela Arkway, [Folk psychological explanation, and causal laws](#)
- Angela Arkway, [The simulation theory and explanations that "make sense of behavior"](#)
- Allison Barnes & Paul Thagard, [Empathy and analogy](#)
- Peter Carruthers, [Autism as mindblindness: an elaboration and partial defence](#)
- Peter Carruthers, [Simulation and self-knowledge: A defence of theory-theory](#)
- Martin Davies and Tony Stone, [Folk psychology and mental simulation](#)
- Martin Davies and Tony Stone, [Mental simulation, tacit theory, and the threat of collapse](#)
- Jim Garson, [Simulation and connectionism: What is the connection?](#)

- Christopher Gauker, [The belief-desire law](#)
- Brie Gertler, [Simulation theory on conceptual grounds](#)
- Alvin Goldman, [The psychology of folk psychology](#)
- Alvin Goldman, [Folk psychology and mental concepts](#)
- Robert Gordon, [Folk psychology as simulation](#)
- Robert Gordon, [Simulation and the explanation of action](#)
- Robert Gordon, [Autism and the "theory of mind" debate](#)
- Robert Gordon, [Sellars's Ryleans revisited](#)
- Robert Gordon, [Developing commonsense psychology: Experimental data and philosophical data](#)
- Robert Gordon, [Reason explanations and counterfactuals](#)
- Robert Gordon & Joe Cruz, [Simulation theory](#)
- Shaun Nichols, Stephen Stich, Alan Leslie & David Klein, [Varieties of off-line simulation](#)
- Shaun Nichols, Stephen Stich, & Alan Leslie, [Choice effects and the ineffectiveness of simulation](#)
- Shaun Nichols & Stephen Stich, [Rethinking co-cognition](#)
- Gerard O'Brien, [A conflation of folk psychologies](#)
- Gloria Origgi, [Theories of theories of mind](#)
- Teed Rockwell, [Eliminative materialism](#)
- Eric Schwitzgebel, [A difficulty for simulation theory due to the close connection of pretense and action in early childhood](#)
- Thomas Suddendorf & Claire Fletcher-Flinn, [Theory of mind and the origins of divergent thinking](#)
- Stephen Stich & Shaun Nichols, [Folk psychology: Simulation or tacit theory?](#)
- Stephen Stich & Shaun Nichols, [Second thoughts on simulation](#)
- Stephen Stich & Shaun Nichols, [Theory theory to the max](#)
- Stephen Stich & Shaun Nichols, [Cognitive penetrability, rationality, and restricted simulation](#)
- Nigel Thomas, [Attitude and image, or, What will simulation theory let us eliminate?](#)
- Dôna Warren, [Externalism and causality: Simulation and the prospects for a reconciliation](#)
- Bill Wilkerson, [Simulation, theory, and the frame problem: The interpretive moment](#)
- Crispin Wright, [What could anti-realism about ordinary psychology possibly be?](#)

Metaphysics of Mind

- William Bechtel and Jennifer Mundale, [Multiple realizability revisited](#)
- Mark Bickhard & Donald Campbell, [Emergence](#)
- John Bickle, [Concepts of intertheoretic reduction in contemporary philosophy of mind](#)
- Ned Block, [Anti-reductionism slaps back](#)
- Ned Block, [Do causal powers drain away?](#)
- Ned Block, [What is functionalism?](#)

- Alex Byrne, [Behaviourism](#)
- Neil Campbell, [Anomalous monism](#)
- Rudolf Carnap, [Psychology in physical language](#)
- Richard Campbell & Mark Bickhard, [Physicalism, emergence, and downward causation](#)
- Daniel Dennett, [Real patterns](#)
- Daniel Dennett, [Two black boxes: A fable](#)
- Daniel Dennett, [The case for rorts](#)
- Crawford Elder, [Mental causation vs. physical causation: No contest](#)
- Crawford Elder, [Physicalism and the fallacy of composition](#)
- Paul Feyerabend, [Mental events and the brain](#)
- Brie Gertler, [Functionalism's methodological predicament](#)
- George Graham, [Behaviorism](#)
- John Heil, [Metaphysics of mind](#)
- Ted Honderich, [Donald Davidson's anomalous monism and the champion of mauve](#)
- Steven Horst, [Mind and the World of Nature](#)
- Dan Hutto, [Davidson's identity crisis](#)
- Henry Jackman, [Belief, rationality, and psychophysical laws](#)
- William Lycan, [A particularly compelling refutation of eliminative materialism](#)
- Ron McClamrock, [Emergence unscathed: Kim on multiple realizability and causal types](#)
- Ruth Millikan, [Historical kinds and the special sciences](#)
- Ruth Millikan, [What is behavior?](#)
- U.T. Place, [Identity theories](#)
- John Post, [Breakwater: The new wave, supervenience and individualism](#)
- Huw Price, [Psychology in perspective](#)
- Hilary Putnam, [The nature of mental states](#)
- Erwin Rogler, [David Lewis's philosophy of mind](#)
- William Seager, [Emergence and efficacy](#)
- William Seager, [Emergence and supervenience](#)
- William Seager, [Generalized epiphenomenalism](#)
- William Seager, [Real patterns and surface metaphysics](#)
- J.J.C. Smart, [The identity theory of mind](#)
- Elliott Sober, [Physicalism from a probabilistic point of view](#)
- Stephen Stich, [Deconstructing the mind](#)
- Stephen Stich and Stephen Laurence, [Intentionality and naturalism](#)
- Daniel Stoljar, [Physicalism](#)
- Tim van Gelder, [Beyond the mind-body problem](#)
- Tim van Gelder, [Monism, dualism, pluralism](#)
- Tim van Gelder, [The distinction between mind and cognition](#)

- J.B. Watson, [The behaviorist's view of thought](#)
- Crispin Wright, [What could anti-realism about ordinary psychology possibly be?](#)

The Self and Personal Identity

- Torin Alter, [Epistemicism and the combined-spectrum argument](#)
- Richard Baron, [The self is unreal](#)
- Lynne Rudder Baker, [Precis of *Persons and Bodies: A Constitution View* \(amd discussion\)](#)
- Tim Bayne, [Persons, animals, and bodies](#)
- Tim Bayne, [Dissociative identity disorder and serious moral status](#)
- Tim Bayne, [The inclusion model of the Incarnation: Problems and prospects](#)
- Jose Luis Bermudez, [Aspects of the self: John Campbell's *Past, Space, and Self*](#)
- Stephen Braude, [Multiple personality and moral responsibility](#) (and [replies](#))
- Stephen Clark, [Minds, memes, and multiples](#) (and [replies](#))
- Daniel Dennett, [The self as a center of narrative gravity](#)
- Daniel Dennett, [The origins of selves](#)
- Daniel Dennett, [Where am I?](#)
- [Shaun Gallagher, Philosophical conceptions of the self](#)
- Grant Gillett, [A discursive account of multiple personality disorder](#) (and Stephen Braude's [reply](#))
- Douglas Hofstadter, [A conversation with Einstein's brain](#)
- Dan Hutto, [Composing our "selves": Aristotelian and fictional personhood](#)
- Dan Hutto, [The story of the self](#)
- Nicholas Humphrey & Daniel Dennett, [Speaking for our selves](#)
- John Locke, [Of identity and diversity](#)
- E.J. Lowe, [Self, agency, and mental causation](#)
- Max More, [The diachronic self](#)
- John Perry, [The self](#)
- Roland Puccetti, [Dennett on the split brain](#) (and [replies](#))
- David Rosenthal, [Emotions and the self](#)
- David Rosenthal, [Persons, minds, and consciousness](#)
- David Sanford, [Where was I?](#)
- William Seager, [The constructed and the secret self](#)
- Ted Sider, [Criteria of personal identity and the limits of conceptual analysis](#)
- Galen Strawson, [The self](#)
- Joe Strout, [Mind uploading home page](#)
- Peter Suber, [Self-determination and selfhood in recent legal cases](#)
- Peter Unger, [The survival of the sentient](#)

Free Will

- Torin Alter & Russell Daw, [Free acts and robot cats](#) (and [reply to Heller](#))
- Derek Farnell, [The free will fiction](#)
- John Martin Fischer & Mark Ravizza, [Morally responsible people without freedom](#)
- Harry Frankfurt, [Alternate possibilities and moral responsibility](#)
- John Fulda, [Partially resolving the tension between omniscience and free will: A mathematical argument](#)
- Patricia Greenspan, [Free will and genetic determinism: Locating the problem\(s\)](#)
- Patricia Greenspan, [Genes, electrotransmitters, and free will](#)
- Thomas Hobbes, [Causation, determinism, and their compatibility with freedom](#)
- Ted Honderich, [Free will, determinism, and moral responsibility: The whole thing in brief](#)
- Ted Honderich, [Compatibilism, incompatibilism, and the smart aleck](#)
- Ted Honderich, [Determinism as true, compatibilism and incompatibilism as both false, and the real problem](#)
- Ted Honderich, [Mind the guff](#)
- Ted Honderich, [After compatibilism and incompatibilism](#)
- David Hume, [The obviousness of determinism](#)
- David Hume, [Freedom reconciled with determinism](#)
- Robert Kane, [Reflections on free will, determinism, and indeterminism](#)
- Immanuel Kant, [For determinism and indeterminism](#)
- Tomis Kapitan, [Modal principles in the metaphysics of free will](#)
- Tomis Kapitan, [Autonomy and manipulated freedom](#)
- Tomis Kapitan, [Agency and omniscience](#)
- Tomis Kapitan, [Deliberation and the presumption of open alternatives](#)
- Ned Markosian, [A compatibilist version of the theory of agent causation](#)
- Thomas Nagel, [Freedom and the view from nowhere](#)
- Robert Pendleton, [Time and free will](#)
- Dirk Pereboom, [Meaning in life without free will](#)
- Aaron Sloman, [How to dispose of the free will issue](#)
- Peter Strawson, [Freedom and resentment](#)
- Christopher Taylor & Daniel Dennett, [Who's afraid of determinism? Rethinking causes and possibilities](#)
- Peter van Inwagen, [The mystery of metaphysical freedom](#)
- J. David Velleman, [Freedom](#) (from [Practical Reflection](#))

Philosophy of Mind (Misc.)

- Colin Allen, [Philosophy of cognitive ethology](#)
- Anthony Atkinson, [Persons, systems, and subsystems](#)

- Andrew Carpenter, [Kant's philosophy of mind](#)
- Daniel Dennett, [Memes and the exploitation of imagination](#)
- Daniel Dennett, [Granny's campaign for safe science](#)
- Daniel Dennett, [Memes: Myths, misunderstandings, and misgivings](#)
- Rick Grush, [Manifolds, co-ordinations, imagination, objectivity](#)
- Colin McGinn, [The problem of philosophy](#)
- Ron McClamrock, [Holism without tears: Local and global effects in cognitive processes](#)
- Ron McClamrock, [Functional analysis and etiology](#)
- Ron McClamrock, [Marr's three levels: A re-evaluation](#)
- Nigel Thomas, [Are theories of imagery theories of imagination?](#)

The Chinese room

- Selmer Bringsjord, [Real robots and the missing thought-experiment in the Chinese-room dialectic](#)
- Steven Brown, [Peirce and formalization of thought: The Chinese room argument](#)
- Jack Copeland, [The curious case of the Chinese gym](#)
- Stevan Harnad, [Minds, machines, and Searle](#)
- Stevan Harnad et al, [Symposium on Symbolism, Connectionism, and the Chinese room](#)
- David Harrison, [Connectionism hits the Chinese gym](#)
- Larry Hauser, [The Chinese room argument](#)
- Larry Hauser, [Searle's Chinese Box: The Chinese Room Argument and Artificial Intelligence](#)
- Larry Hauser, [Searle's Chinese box: Debunking the Chinese room argument](#)
- Larry Hauser, [Nixin' goes to China](#)
- Pat Hayes/Stevan Harnad/Don Perlis/Ned Block, [Virtual symposium on virtual mind](#)
- Diane Law, [Searle, subsymbolic functionalism, and synthetic intelligence](#)
- William Rapaport, [Understanding understanding: Syntactic semantics and computational cognition](#)
- John Searle, [Minds, brains, and programs](#)
- John Searle, [Is the brain a digital computer?](#)
- John Y.R. Shen, [A real Chinese room](#)

Gödel's theorem and AI

- Damjan Bojadziev, [Gödel's theorem for minds and computers](#)
- Selmer Bringsjord, [A refutation of Penrose's new Gödelian case against the computational conception of mind](#)
- David Chalmers, [Minds, machines, and mathematics](#)
- Kari Coleman, [Gödel propositions for the mind](#)
- Jack Copeland, [Turing's O-machines, Searle, Penrose, and the brain](#)

- Daniel Dennett, [Review of *The Emperor's New Mind*](#)
- Harry Deutsch, [Deconstructing mathematics and mind: Some implications of Gödel's incompleteness theorems](#)
- Solomon Feferman, [Penrose's Gödelian argument](#)
- Rick Grush & Patricia Churchland, [Gaps in Penrose's toilings](#)
- Geoffrey Laforte, Pat Hayes, & Ken Ford, [Why Godel's theorem cannot refute computationalism: A reply to Penrose](#)
- J. R. Lucas, [Minds, machines, and Gödel](#)
- J. R. Lucas, [Satan stultified: A rejoinder to Paul Benacerraf](#)
- J. R. Lucas, [The Gödelian argument: Turn over the page](#)
- J. R. Lucas, [The implications of Gödel's theorem](#)
- Tim Maudlin, [Between the motion and the act...](#)
- Daryl McCullough, [Can humans escape Gödel?](#)
- Drew McDermott, [\[STAR\] Penrose is wrong](#)
- Roger Penrose, [Beyond the doubting of a shadow](#)
- Hilary Putnam, [Review of *Shadows of the Mind*](#)
- Brian Rosmaita, [Minds, machines, and metamathematics](#)
- William Seager, [Yesterday's algorithm: Penrose and the Godel argument](#)
- Aaron Sloman, [The emperor's real mind](#)

The Turing Test

- Selmer Bringsjord, [Animals, zombanimals, and the total Turing test: The essence of artificial intelligence](#)
- Selmer Bringsjord, [Creativity, the Turing test, and the \(better\) Lovelace test](#)
- Bruce Edmonds, [The constructability of artificial intelligence \(as defined by the Turing test\)](#)
- Robert French, [Subcognition and the limits of the Turing test](#)
- Robert French, [Refocusing the debate on the Turing Test: A reply to Jacquette](#)
- Stevan Harnad, [The Turing test is not a trick](#)
- Stevan Harnad, [Turing on reverse-engineering the mind](#)
- Stevan Harnad, [Other bodies, other minds](#) (and Larry Hauser's [reply](#))
- Jason Hutchens, [How to pass the Turing test by cheating](#)
- William Rapaport, [How to pass a Turing test: Syntactic semantics, natural-language understanding, and first-person cognition](#)
- James Sennett, [The ice man cometh: Lt. Commander Data and the Turing test](#)
- Stuart Shieber, [Lessons from a restricted Turing test](#) (and Hugh Loebner's [reply](#))
- Alan Turing, [Computing machinery and intelligence](#)
- Peter Turney, [Answering subcognitive Turing Test questions: A reply to French](#)

- Stuart Watt, [Naive psychology and the inverted Turing test](#) (and [comments](#))
- Blay Whitby, [Why the Turing test is AI's biggest blind alley](#)

Philosophy of Connectionism

- Murat Aydede, [Language of thought: The connectionist contribution](#)
- Murat Aydede, [Connectionism and the language of thought](#)
- William Bechtel, [Embodied connectionism](#)
- William Bechtel, [What should a connectionist philosophy of science look like?](#)
- William Bechtel, [Natural deduction in connectionist systems](#)
- István Berkeley, [Connectionism reconsidered: Minds, machines, and models](#)
- István Berkeley, [What the #*\\$%! is a subsymbol?](#)
- István Berkeley, [Some myths of connectionism](#)
- David Chalmers, [Syntactic transformations on distributed representations](#)
- David Chalmers, [Connectionism and compositionality: Why Fodor and Pylyshyn were wrong](#)
- David Chalmers, [The evolution of learning: An experiment in genetic connectionism](#)
- Andy Clark & Chris Eliasmith, [Philosophical issues in brain theory and connectionism](#)
- Jack Copeland, [On Alan Turing's anticipation of connectionism](#)
- Daniel Dennett, [Mother nature vs the walking encyclopedia](#)
- Chris Eliasmith, [Structure without symbols: Providing a distributed account of high-level cognition](#)
- Eric Lormand, [Classical and connectionist models](#)
- Eric Lormand, [Connectionist languages of thought](#)
- Eric Lormand, [Connectionist content](#)
- L. Niklasson & T. van Gelder, [Can connectionist models exhibit non-classical structure sensitivity?](#)
- Gerard O'Brien, [Is connectionism common sense?](#)
- Gerard O'Brien, [The role of implementation in connectionist explanation.](#)
- Gerard O'Brien, [Connectionism, analogicity and mental content.](#)
- Tim van Gelder & L. Niklasson, [On being systematically connectionist](#)
- Tim van Gelder, [Connectionism and the mind-body problem: Exposing the distinction between mind and cognition](#)
- Tim van Gelder, [Why distributed representation is inherently non-symbolic](#)
- J. Waskan & W. Bechtel, [Directions in connectionist research: Tractable computations without syntactically structured representations](#)

Computationalism

- Mark Bickhard, [The biological foundations of cognitive science](#)

- Ned Block, [The mind as the software of the brain](#)
- Selmer Bringsjord, [Cognition is not computation: The argument from irreversibility](#)
- David Buller, [Confirmation and the computational paradigm \(or: Why do you think they call it artificial intelligence?\)](#)
- David Chalmers, [A computational foundation for the study of cognition](#)
- David Chalmers, [Does a rock implement every finite-state automaton?](#)
- Robert Cummins, [The mind as computer](#)
- Hubert Dreyfus, [From Socrates to expert systems: The limits and dangers of calculative rationality](#)
- Chris Eliasmith, [Computation and dynamic models of the mind](#)
- Chris Eliasmith, [The third contender: A critical examination of the dynamicist theory of cognition](#)
- Christopher Green, [Is AI the right method for cognitive science?](#) (and [commentaries](#))
- Rick Grush, [The semantic challenge to cognitive neuroscience](#)
- Rick Grush & Patricia Churchland, [Computation and the brain](#)
- Stevan Harnad, [The symbol grounding problem](#)
- Stevan Harnad, [Computation is just interpretable symbol manipulation; Cognition isn't](#)
- Steven Horst, [Symbols and computation: A critique of the computational theory of mind](#)
- William Rapaport, [How minds can be computational systems](#)
- Aaron Sloman, [What are virtual machines? Are they real?](#)
- Aaron Sloman, [Architecture-based conceptions of mind](#)
- Aaron Sloman, [What sort of architecture is required for a human-like agent?](#)
- Tim van Gelder, [The dynamical hypothesis in cognitive science](#) (and [response](#))
- Tim van Gelder, [Defending the dynamic hypothesis](#)
- Tim van Gelder, [Computers and computation in cognitive science](#)

Philosophy of AI, Misc

- John Beloff, [Minds or machines](#)
- Nick Bostrom, [How long before superintelligence?](#)
- Nick Bostrom, [The transhumanist FAQ](#)
- Selmer Bringsjord, [Precis of *What Robots Can and Can't be*](#) (and [commentaries](#))
- William Clancey, [How situated cognition is different from situated robotics](#)
- Daniel Dennett, [Can machines think?](#)
- Daniel Dennett, [Cog as a thought experiment](#)
- Daniel Dennett, [Did HAL committ murder?](#)
- Daniel Dennett, [Cognitive wheels: The frame problem of AI](#)
- Eric Drexler, [Thinking machines](#)
- Matthew Elton, [Robots and rights: The ethical demands of artificial agents](#)
- Larry Hauser, [Why isn't my pocket calculator a thinking thing?](#)

- Jaron Lanier, [Agents of alienation](#)
 - Jaron Lanier, [Mindless thought experiments \(a critique of machine intelligence\)](#)
 - John McCarthy, [Ascribing mental qualities to machines](#)
 - John McCarthy, [What is artificial intelligence?](#)
 - John McCarthy & Patrick Hayes, [Some philosophical problems from the standpoint of artificial intelligence](#)
 - Drew McDermott, [How intelligent is Deep Blue?](#)
 - Marvin Minsky, [Why people think computers can't](#)
 - Aaron Sloman, [The Computer Revolution in Philosophy](#)
 - Aaron Sloman, [The irrelevance of Turing machines to AI](#)
 - Stanford Humanities Review, [Constructions of the mind: Artificial intelligence and the humanities](#)
 - Tim van Gelder, [Into the Deep Blue yonder](#)
 - Vernor Vinge, [The technological singularity](#)
 - Mitchell Waldrop, [Can computers think?](#)
 - Eliezer Yudkowsky, [Creating friendly AI](#)
 - Eliezer Yudkowsky, [General intelligence and seed AI](#)
 - Eliezer Yudkowsky, [Staring into the singularity](#)
-

Go to:

- [Online papers on consciousness](#)
- [David Chalmers' home page](#)

Online papers on consciousness, part 3: Science of consciousness

Compiled by [David Chalmers](#)

Part 3 of [Online papers on consciousness](#)

- [Part 1: Philosophy of consciousness](#)
- [Part 2: Other philosophy of mind](#)
- Part 3: Science of Consciousness

TABLE OF CONTENTS [329 papers]

- [Neuroscience of visual consciousness](#) [23]
- [Consciousness and neuroscience](#) [34]
- [Cognitive models of consciousness](#) [32]
- [Unconscious perception](#) [18]
- [Implicit memory](#) [15]
- [Implicit learning](#) [10]
- [Change blindness and inattention blindness](#) [14]
- [Visual consciousness \(misc\)](#) [8]
- [Consciousness and psychology \(misc.\)](#) [23]
- [Consciousness in the history of psychology](#) [30]
- [Consciousness and time](#) [7]
- [Animal consciousness](#) [13]
- [Consciousness and artificial intelligence](#) [11]
- [Consciousness and physics](#) [37]
- [Phenomenology](#) [37]
- [Miscellaneous](#) [19]

Neuroscience of Visual Consciousness

- Paul Azzopardi & Alan Cowey, [Is blindsight like normal, near-threshold vision?](#)
- Moshe Bar & Irving Biederman, [Localizing the cortical region mediating visual awareness of object identity](#)
- Andreas Engel, Pascal Fries, Pieter Roelfsema, Peter König, & Wolf Singer, [Temporal binding, binocular rivalry, and consciousness](#) (and [discussion](#))
- Pascal Fries, Pieter Roelfsema, Andreas Engel, Peter König, & Wolf Singer, [Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry](#)
- Charles M. Gray & Gonzalo Viana Di Prisco, [Stimulus-dependent neuronal oscillations and local](#)

[synchronization in striate cortex of the alert cat](#)

- Sabine Kastner & Leslie Ungerleider, [Mechanisms of visual attention in the human cortex](#)
- Christof Koch, [Towards the neuronal substrate of visual consciousness](#)
- Ilona Kovacs, Thomas Papathomas, Ming Yang, & Akos Feher, [When the brain changes its mind: Interocular grouping during binocular rivalry](#)
- Stephen Luck, Leonardo Chelazzi, Steven Hillyard, & Robert Desimone, [Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex](#)
- Steven Luck & Michelle Ford, [On the role of selective attention in visual perception](#)
- Erik Lumer & Geraint Rees, [Covariation of activity in visual and prefrontal cortex associated with subjective visual perception](#)
- Erik Lumer, K.J. Friston, & Geraint Rees, [Neural correlates of perceptual rivalry in the human brain](#)
- David Milner and Melvyn Goodale, [The visual brain in action \(precis\)](#), and [commentaries](#)
- Stephen L. Macknik & Michael M. Haglund, [Optical images of visible and invisible percepts in the primary visual cortex of primates](#)
- A. Sahraie, L. Weiskrantz, J.L. Barbur, et al, [Pattern of neuronal activity associated with conscious and unconscious processing of visual signals](#)
- D.L. Sheinberg & N. Logothetis, [The role of temporal cortical areas in perceptual organization](#)
- John Skoyles, [Another variety of vision](#)
- R. Srinivasan, D.P. Russell, G.M. Edelman, & G. Tononi, [Increased synchronization of neuromagnetic responses during conscious perception"](#)
- G. Tononi, R. Srinivasan, D.P. Russell, & G.M. Edelman, [Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses](#)
- Marius Usher & Nick Donnelly, [Visual synchrony affects binding and segmentation in perception](#)
- S. Zeki, S. Aglioti, D. McKeefry, & G. Berlucchi, [The neurological basis of conscious color perception in a blind patient](#)

Consciousness and Neuroscience, Misc

- Giorgio Ascoli, [The complex link between neuroanatomy and consciousness](#)
- Bernard Baars, [The neural basis of conscious experience](#)
- Ned Block, [How to find the neural correlate of consciousness](#)
- Massimo Bondi & Manuele Bondi, [The role of synaptic junctions in the identification of human consciousness](#)
- David Chalmers, [On the search for the neural correlate of consciousness](#)
- David Chalmers, [What is a neural correlate of consciousness?](#)
- Patricia Churchland, [Can neurobiology teach us anything about consciousness?](#)
- William Clancey, [The biology of consciousness: Comparative review of Rosenfield and Edelman](#)

- Axel Cleeremans & John-Dylan Haynes, [Correlating consciousness: A view from empirical science](#)
- Rodney Cotterill, [On the mechanism of consciousness](#)
- Rodney Cotterill, [On the neural correlates of consciousness](#)
- Rodney Cotterill, [Prediction and internal feedback in conscious perception](#)
- Francis Crick and Christof Koch, [Consciousness and neuroscience](#)
- Francis Crick and Christof Koch, [The problem of consciousness](#)
- Francis Crick and Christof Koch, [Why neuroscience may be able to explain consciousness](#)
- Francis Crick and Christof Koch, [The unconscious homunculus](#)
- John C. Eccles, [The effect of silent thinking on the cerebral cortex](#)
- Jeffrey Gray, [The contents of consciousness: A neuropsychological conjecture](#)
- Susan Greenfield, [Neural assemblies and consciousness](#)
- Rudolf Hernegger, [Changes of paradigm in consciousness research](#)
- J. Allan Hobson, Edward Pace-Schott, & Robert Stickgold, [Dreaming and the brain: Toward a cognitive neuroscience of conscious states](#)
- Fahmeed Hyder, Elizabeth Phelps, Robert Shulman, et al, ["Willed action": A functional MRI study of the human prefrontal cortex during a sensorimotor task](#)
- Stephen Jones, [Introduction to the physiology of ordinary consciousness](#)
- Alain Morin, [Right hemispheric self-awareness](#)
- Alain Morin, [The split-brain debate revisited: On the importance of language and self-recognition for right hemispheric consciousness](#)
- Alain Morin, [On a relation between self-awareness and inner speech: Additional evidence from brain studies](#)
- James Newman, [Thalamocortical foundations of conscious experience](#) (and [discussion](#))
- Alva Noe & Evan Thompson, [Neural correlates of consciousness and the matching content doctrine](#)
- C. Portas, G. Rees, A Howseman, O. Josephs, R. Turner, & C.D. Frith, [A specific role for the thalamus in mediating the interaction of attention and arousal in humans](#)
- John Smythies, [The biochemical basis of coma](#)
- Sean Spence, [Free will in the light of neuropsychiatry](#) (and [replies](#))
- Ruediger Vaas, [Why neural correlates of consciousness are fine, but not enough](#)
- Douglas Watt, [Emotion and consciousness: Implications of affective neuroscience for ERTAS theories of consciousness](#)

Cognitive Models of Consciousness

- Bernard Baars, [A Cognitive Theory of Consciousness](#)
- Bernard Baars, [In the theatre of consciousness](#)
- Myles Bogner, Uma Ramamurthy, & Stan Franklin, [Consciousness and conceptual learning in a](#)

[socially situated agent](#)

- Bruce Bridgeman, [On the evolution of consciousness and language](#) (and [replies](#))
- William H. Calvin, [Competing for consciousness: A Darwinian mechanism at an appropriate level of explanation](#)
- Fu Chang, [A theory of consciousness](#)
- Rodney Cotterill, [Navigation, consciousness, and the body/mind problem](#)
- Rodney Cotterill, [On the unity of conscious experience](#)
- Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux, [A neuronal model of a global workspace in effortful cognitive tasks](#)
- Philip Dorrell, [Computation vs. feelings and the production/judgment model](#)
- Peter Ells, [The decider system model: A defence of the Cartesian theatre](#)
- Stan Franklin, [Conscious software: A computational view of mind](#)
- Stan Franklin, [Action selection and language generation in "conscious" software agents](#)
- Stan Franklin and Art Graesser, [A software agent model of consciousness](#)
- Stevan Harnad, [Consciousness: An afterthought](#)
- Anthony Jack & Tim Shallice, [Introspective physicalism as an approach to the science of consciousness](#)
- Steven Lehar, [Gestalt isomorphism and the primacy of the subjective conscious experience: A gestalt bubble model](#)
- Don Mathis & Michael Mozer, [On the computational utility of consciousness](#)
- Josh McDermott, [Global workspace theory: Consciousness explained?](#)
- George McKee, [The engine of awareness: Autonomous synchronous representations](#)
- Thomas Metzinger, [Faster than thought: Holism, homogeneity, and temporal coding](#)(and [comments](#))
- Erik Mueller & Michael Dyer, [Towards a computational theory of human daydreaming](#)
- Shaun Nichols & Stephen Stich, [Reading one's own mind: A cognitive theory of self-awareness](#)
- Gerard O'Brien & Jon Opie, [Cognitive science and phenomenal consciousness: A dilemma, and how to avoid it](#)
- Gerard O'Brien & Jon Opie, [A connectionist theory of phenomenal experience](#)
- Ruadhan O'Flanagan, [Underlying mechanisms of consciousness](#)
- John Taylor, [Modeling consciousness](#)
- John Taylor, [The relational mind](#)
- John Taylor, [Constructing the relational mind](#)
- Christoph von der Malsburg, [The coherence definition of consciousness](#)

Unconscious Perception

- Barry Beyerstein & Eric Eich, [Subliminal self-help tapes: Promises, promises...](#)

- Robert Dell'Acqua & Jonathan Grainger, [Unconscious semantic priming from pictures](#)
- Sean Draine & Anthony Greenwald, [Replicable unconscious semantic priming](#) (and Reingold and Merikle's [commentary](#))
- Ken Forster, [The pros and cons of masked priming](#)
- S. Kouider & E. Dupoux, [A functional disconnection between spoken and visual word recognition: Evidence from unconscious priming](#)
- Pawel Lewicki, Thomas Hill, & Maria Czyzewska, [Nonconscious acquisition of information](#)
- Don Mathis & Michael Mozer, [Conscious and unconscious perception: A computational theory](#)
- Philip Merikle & Meredyth Daneman, [Psychological investigations of unconscious perception](#)
- Philip Merikle & Meredyth Daneman, [Memory for unconsciously perceived events: Evidence from anesthetized patients](#)
- Philip Merikle & Eyal Reingold, [Recognition and lexical decision without detection: Unconscious perception?](#)
- Philip Merikle & Eyal Reingold, [Measuring unconscious perceptual processes](#)
- Timothy Moore, [Subliminal self-help auditory tapes: An empirical test of perceptual consequences](#)
- M.J. Morgan, A.J.S. Mason, & J.A. Solomon, ["Blindsight" in normal subjects?](#)
- Eyal Reingold & Philip Merikle, [Using direct and indirect measures to study perception without awareness](#)
- Eyal Reingold & Philip Merikle, [In the inter-related of theory and measurement in the study of unconscious processes](#)
- John Taylor, [Breakthrough to awareness](#)
- Max Velmans, [When perception becomes conscious](#)

Implicit Memory

- Sean Draine, Anthony Greenwald, & Mahzarin Banaji, [Modeling unconscious gender bias in fame judgments.](#)
- Emrah Duzel, Andrew Yonelinas, Endel Tulving, et al, [Event-related brain potential correlates of two states of conscious awareness in memory](#)
- Andrew Mayes, Patricia Gooding, & Rob van Eijk, [A new theoretical framework for explicit and implicit memory](#)
- Gail McKoon & Roger Ratcliff, [How should implicit memory phenomena be modeled?](#)
- Philip Merikle & Eyal Reingold, [Comparing direct \(explicit\) to indirect \(implicit\) measures to study unconscious memory](#)
- Roger Ratcliff & Gail McKoon, [Bias effects in implicit memory tasks](#)
- Roger Ratcliff, Trish Van Zandt, & Gail McKoon, [Two factor theory, single process theories, and recognition memory](#)
- Eyal Reingold & Jeffrey Toth, [Process dissociations versus task dissociations: A controversy in](#)

[progress](#)

- Daniel Schacter, [Illusory memories: A cognitive neuroscience analysis](#)
- Larry Squire & Stuart Zola, [Structure and function of declarative and nondeclarative memory systems](#)
- Patrice Terrier, [Re-examining the role of consistency](#)
- R.F. Thompson & J.J. Kim, [Memory systems in the brain and the localization of a memory](#)
- Jeffrey Toth, Eyal Reingold, & Larry Jacoby, [Toward a redefinition of implicit memory: Process dissociations following elaborative processing and self-generation](#)
- Jeffrey Toth, Eyal Reingold, & Larry Jacoby, [A response to Graf and Komamatsu's critique of the process-dissociation procedure: When is caution necessary?](#)
- Daniel Willingham & Laura Preuss, [The death of implicit memory](#)

Implicit Learning

- Maud Boyer, Arnaud Destrebecqz, and Axel Cleeremans, [The serial reaction task: Learning without knowing, or knowing without learning?](#)
- Axel Cleeremans, [Principles for implicit learning](#)
- Axel Cleeremans, [Implicit sequence learning: The truth is in the details](#)
- Axel Cleeremans, Arnaud Destrebecqz, & Maud Boyer, [Implicit learning: News from the front](#)
- Tim Curran, [On the neural mechanisms of sequence learning](#)
- Zoltan Dienes & Josef Perner, [A theory of implicit and explicit knowledge](#)
- Georgina Jackson & Stephen Jackson, [Do measures of explicit learning actually measure what is being learnt in the serial reaction time task?](#) (and [commentary](#))
- Luis Jiménez, Castor Méndez, & Axel Cleeremans, [Comparing direct and indirect measures of implicit learning](#)
- David Shanks & Mark St. John, [Characteristics of dissociable human learning systems](#)

Change Blindness and Inattentional Blindness

- Jochen Braun, [Inattentional blindness: It's great but not necessarily about attention](#)
- Donelson Dulany, [Inattentional awareness](#)
- Arien Mack and Irvin Rock, [Inattentional blindness: An overview](#)
- S. Most, D. Simons, B. Scholl, & C. Chabris, [Sustained inattentional blindness: The role of location in the detection of unexpected dynamic events](#)
- Cathleen Moore, [Inattentional blindness: Perception or memory and what does it matter?](#)
- Alva Noe and Evan Thompson, [Beyond the grand illusion: What change blindness really teaches us about vision](#)
- Alva Noe and Kevin O'Regan, [Perception, attention, and the grand illusion](#)
- Kevin O'Regan, H. Deubel, J.J. Clark, & R.A. Rensink, [Picture changes during blinks: Looking](#)

[without seeing and seeing without looking](#)

- Kevin O'Regan, Ronald Rensink, & James Clark, [Change blindness as a result of mudsplashes](#)
- G. Rees, C. Russell, C.D. Frith, & J. Driver, [Inattention blindness versus inattentional amnesia for fixated but ignored words](#)
- Ronald Rensink, [Seeing, sensing, and scrutinizing](#)
- Ronald Rensink, [When good observers go bad: Change blindness, inattentional blindness, and visual experience](#)
- Patrick Wilken, Kevin Korb, William Webster, and Damian Conway, [How many items can we keep in consciousness? Experiments in visual short-term memory and change blindness](#)
- Jeremy Wolfe, [Inattentional amnesia](#)

Visual Consciousness (Misc.)

- Paul Bach-y-Rita, [Sensory substitution and qualia](#)
- Paul Bach-y-Rita, Mitchell Tyler, & Kurt Kaczmarek, [Seeing with the brain](#)
- Daniel Dennett, [Filling in vs. finding out: A ubiquitous confusion in cognitive science](#)
- Lianggang Lou, [Selective peripheral fading: Evidence for inhibitory effect of attention on visual sensation](#)
- Kevin O'Regan, [Solving the "real" mysteries of visual perception: The world as an outside memory](#)
- Kevin O'Regan & Alva Noe, [A sensorimotor account of vision and visual consciousness](#)
- Luiz Pessoa, Evan Thompson, & Alva Noe, [Finding out about filling in: A guide to perceptual completion for visual science and the philosophy of perception](#)
- Daniel Richardson & Michael Spivey, [Representation, space and Hollywood Squares: Looking at things that aren't there anymore](#)

Consciousness and Psychology (Misc.)

- Jorge Avelaira, [Consciousness and reality: A stable-dynamic model based on Jungian psychology](#)
- Bernard Baars, [Understanding subjectivity: Global workspace theory and the resurrection of the observing self](#) (and [comments](#))
- Bernard Baars, [A thoroughly empirical approach to consciousness](#) (and [replies](#))
- Bernard Baars, [Metaphors of consciousness and attention in the brain](#) (and [discussion](#))
- Rodney Cotterill, [Autism, intelligence, and consciousness](#)
- Richard Cytowic, [Synesthesia: Phenomenology and neuropsychology](#) (and [comments](#))
- Valerie Hardcastle, [Locating Consciousness \(precis\)](#) (and [comments](#))
- Valerie Hardcastle, [Attention versus consciousness: A distinction with a difference](#)
- James Hopkins, [The unconscious](#)
- Steven Horst, [Phenomenology and psychophysics](#)

- David LaBerge, [Defining awareness by the triangular circuit of attention](#) (and [comments](#))
- Stephen LaBerge, [Lucid dreaming: Psychophysiological studies of consciousness during REM sleep](#)
- George Mandler, [Consciousness redux](#)
- Stephen Palmer, [Color, consciousness, and the isomorphism constraint](#)
- David Rosenthal, [Consciousness and metacognition](#)
- Marks Solms, [What is consciousness?](#) (and [replies](#))

Consciousness in the History of Psychology

- Bill Adams, [Why the introspectionist school went bankrupt](#)
- James Rowland Angell, [The province of functional psychology](#)
- James Mark Baldwin, [Consciousness and evolution](#)
- Knight Dunlap, [The case against introspection](#)
- Gustav Fechner, [The measurement of sensation](#)
- Sigmund Freud, [The structure of the unconscious](#)
- Willard Gore, [Image or sensation?](#)
- Hermann Helmholtz, [The facts of perception](#)
- William James, [The stream of consciousness](#)
- Kurt Koffka, [Perception: An introduction to the gestalt theory](#) (with Christopher Green's [introduction](#))
- Wolfgang Kohler, [Gestalt psychology today](#)
- Wolfgang Kohler, [An old pseudoproblem](#)
- George Trumbull Ladd, [Consciousness and evolution](#)
- Karl Lashley, [The behavioristic interpretation of consciousness](#)
- Wilfrid Lay, [Organic images](#)
- Ernst Mach, [The analysis of sensations](#)
- Teed Rockwell, [The effects of atomistic ontology on the history of psychology](#)
- Eugene Taylor & Robert Wozniak, [Pure experience: The response to William James](#)
- Edward L. Thorndike, [The study of consciousness and the study of behavior](#)
- E.B. Titchener, [Brentano and Wundt: Empirical and experimental psychology](#)
- E.B. Titchener, [The postulates of a structural psychology](#)
- E.B. Titchener, [The schema of introspection](#)
- Margaret Washburn, [Introspection as an objective method](#)
- John B. Watson, [Psychology as the behaviorist views it](#) (and Titchener's [comment](#))
- Max Wertheimer, [Gestalt theory](#)
- Max Wertheimer, [Laws of organization in perceptual forms](#)
- Robert H. Wozniak, [Mind and Body: Rene Descartes to William James](#)

- Wilhelm Wundt, [Outlines of Psychology](#)

Consciousness and Time

- Daniel Dennett & Marcel Kinsbourne, [Time and the observer: The where and when of consciousness in the brain](#) (and [reply to commentators](#))
- Celia Green & Grant Gillett, [Are mental events preceded by their physical causes?](#)
- Ted Honderich, [Is the mind ahead of the brain? Benjamin Libet's evidence examined](#)
- Ted Honderich, [Is the mind ahead of the brain? Rejoinder to Benjamin Libet](#)
- Francisco Varela, [The specious present: A neurophenomenology of time consciousness](#)
- Fred Alan Wolf, [The timing of conscious experience: A causality-violating interpretation](#)

Animal Consciousness

- Colin Allen, [Animal consciousness](#)
- Colin Allen, [Consciousness: Essential or dispensable?](#)
- Colin Allen & Marc Bekoff, [Species of Mind](#)
- Peter Carruthers, [Animal subjectivity](#) (and [commentaries](#))
- Peter Carruthers, [Sympathy and subjectivity](#)
- Daniel Dennett, [Animal consciousness: What matters and why](#)
- Matthew Elton, [Human and animal consciousness](#)
- Gordon Gallup, [Are animals self-aware? Yes](#)
- Thomas Huxley, [On the hypothesis that animals are automata](#)
- C. Lloyd Morgan, [An Introduction to Comparative Psychology](#)
- Daniel Povinelli, [Are animals self-aware? Maybe not](#)
- Peter Singer, [Do animals feel pain?](#)

Consciousness and Artificial Intelligence

- Murat Aydede & Guven Guzeldere, [Consciousness, intentionality, and intelligence: Some foundational issues for artificial intelligence](#)
- William Bechtel, [Consciousness: Perspectives from symbolic and connectionist AI](#)
- Selmer Bringsjord, [What robots can and can't be](#) (and [replies](#))
- Scott Brockmaier, [Computational architecture and the creation of consciousness](#)
- Daniel Dennett, [The practical requirements for making a conscious robot](#)
- Daniel Dennett, [Consciousness in human and robot minds](#)
- Inman Harvey, [Evolving robot consciousness: The easy problems and the rest](#)
- John McCarthy, [Making robots conscious of their mental states](#)
- Marvin Minsky, [Conscious machines](#)

- Mark Sharlow, [Can machines have first-person properties?](#)
- Aaron Sloman, [A systems approach to consciousness](#)

Consciousness and Physics

- Diederik Aerts, Jan Broekaert, & Liane Gabora, [Intrinsic contextuality as the crux of consciousness](#)
- Bernard Baars, [Can physics provide a theory of consciousness?](#)
- Erhard Bieberich, [In search of a neuronal substrate of the human mind: New concepts from "topological neurochemistry"](#)
- Erhard Bieberich, [Structure in human consciousness: A fractal approach to the topology of the self perceiving an outer world in an inner space](#)
- David Bohm, [A new theory of the relationship between mind and matter](#)
- Jeremy Butterfield, [Quantum curiosities of psychophysics](#)
- Alex Byrne & Ned Hall, [Chalmers on consciousness and quantum mechanics](#)
- C. J. S. Clarke, [The nonlocality of mind](#)
- Brian Flanagan, [Are perceptual fields quantum fields?](#)
- Gordon Globus, [Quantum consciousness is cybernetic](#)
- Amit Goswami, [The hard questions: View from a science of consciousness](#)
- Stuart Hameroff, ["Funda-mentality": Is the conscious mind subtly linked to a basic level of the universe?](#)
- Stuart Hameroff, [Did consciousness cause the Cambrian evolutionary explosion?](#)
- Stuart Hameroff, [More neural than thou \(reply to Churchland\)](#)
- Stuart Hameroff, [Quantum computing in microtubules: An intra-neural correlate of consciousness?](#)
- Stuart Hameroff & Roger Penrose, [Conscious events as orchestrated space-time selections](#)
- Stuart Hameroff & Roger Penrose, [Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness](#)
- Stuart Hameroff & Alwyn Scott, [A Sonoran afternoon: Dialogue on quantum mechanics and consciousness](#)
- Stanley Klein, [Is quantum mechanics relevant to understanding consciousness?](#)
- Kirk Ludwig, [Why the difference between quantum and classical mechanics is irrelevant to the mind-body problem](#)
- Jacques Mallah, [The observer as a classical computer in a quantum universe](#)
- John McCrone, [The theory vacuum: On "nutty quantum approaches" to the mind](#)
- Ulrich Mohrhoff, [Quantum mechanics and consciousness: Fact and fiction](#)
- Gregory Mulhauser, [On the end of a quantum mechanical romance](#)
- Don Page, [Attaching theories of consciousness to Bohmian quantum mechanics](#)
- Don Page, [Information loss in black holes and/or conscious beings?](#)

- Don Page, [Mindless sensationalism: A quantum framework for consciousness](#)
- Don Page, [Sensible quantum mechanics: Are probabilities only in the mind?](#)
- Alfredo Pereira, [The quantum mind/classical brain problem](#)
- Osvaldo Pessoa, [What is an essentially quantum mechanical effect?](#)
- Henry Stapp, [Chance, choice, and consciousness: A causal quantum theory of the mind/brain](#)
- Henry Stapp, [Science of consciousness and the hard problem](#)
- Henry Stapp, [The hard problem: A quantum approach](#)
- Henry Stapp, [Why classical mechanics cannot naturally accommodate consciousness but quantum mechanics can](#)
- Henry Stapp, [Whiteheadian process and the quantum theory of mind](#)
- Victor Stenger, [The myth of quantum consciousness](#)
- Ruediger Vaas, [Why quantum correlates of consciousness are fine, but not enough](#)

Phenomenology

- Henk Barendregt, [Buddhist phenomenology](#) (and [part 2](#))
- Glenn Braddock, [Beyond reflection in naturalized phenomenology](#)
- Glenn Braddock, [Perception and judgment in Dennett and Merleau-Ponty](#)
- Steven Brown, [Beyond the fringe: James, Gurwitsch, and the conscious horizon](#)
- Steven Brown, [Tip-of-the-tongue phenomena: An introductory phenomenological analysis](#)
- David Chalmers, [First-person methods in the science of consciousness](#)
- Jonathan Cole, Natalie Depraz, & Shaun Gallagher, [Unity and disunity in bodily awareness: Phenomenology and neuroscience](#)
- Natalie Depraz, Francisco Varela, and Pierre Vermersch, [The gesture of awareness: An account of its structural dynamics](#)
- Daniel Dennett, [The fantasy of first-person science](#)
- Hubert Dreyfus, [The primacy of phenomenology over logical analysis: A critique of Searle](#)
- Hubert Dreyfus, [The current relevance of Merleau-Ponty's phenomenology of embodiment](#)
- Hubert Dreyfus, [Intelligence without representation: Merleau-Ponty's phenomenology and scientific explanation](#) (and [commentaries](#))
- Robert Forman, [What does mysticism have to teach us about consciousness?](#)
- Shaun Gallagher, [Concepts of embodiment in phenomenology and cognitive science](#)
- Shaun Gallagher, [Mutual enlightenment: Recent phenomenology in cognitive science](#)
- Edmund Husserl, [The way into phenomenological transcendental philosophy from psychology](#)
- Edmund Husserl, [Pure phenomenology, its method, and its field of investigation](#)
- Piet Hut, [Exploring actuality through experiment and experience](#)
- Anthony Jack and Tim Shallice, [Introspective physicalism as an approach to the science of consciousness](#)

- Bruce MacLennan, [The investigation of consciousness through phenomenology and neuroscience](#)
- Maurice Merleau-Ponty, [The structure of behavior](#)
- Jacob Needleman, [Inner empiricism as a way to a science of consciousness](#)
- Greg Nixon, [A hermeneutic objection: Language and the inner view](#)
- Morten Overgaard, [The role of phenomenological reports in experiments on consciousness](#)
- Michael Polanyi, [The structure of consciousness](#)
- John Searle, [Limits of phenomenology](#)
- Charles Tart, [Consciousness: A psychological, transpersonal, and parapsychological approach](#)
- Evan Thompson, [Empathy and consciousness](#)
- Evan Thompson, Alva Noe, & Luiz Pessoa, [Perceptual completion: A case study in phenomenology and cognitive science](#)
- Tim van Gelder, [Wooden iron? Husserlian phenomenology meets cognitive science](#)
- Francisco Varela, [Neurophenomenology: A methodological remedy for the hard problem](#)
- Francisco Varela & Jonathan Shear, [First-person methodologies: What, why, how?](#)
- Max Velmans, [A reflexive science of consciousness](#)
- Max Velmans, [Intersubjective science](#)
- Max Velmans, [Heterophenomenology vs. critical phenomenology: A dialogue with Dan Dennett](#)

Miscellaneous

- Bill Adams, [Consciousness on its own terms](#)
- William Calvin, [The Cerebral Code](#)
- William Calvin, [The Cerebral Symphony: Seashore Reflections on the Structure of Consciousness](#)
- Keith Chandler, [The Mind Paradigm](#)
- Ellis Cooper, [Dictionary of consciousness](#)
- Walter Freeman, [Commentary on "The mystery of consciousness"](#)
- Stephen Jones, [The Brain Project](#)
- Brian Josephson & Beverly Rubik, [The challenge of consciousness research](#)
- Jaron Lanier, [Death: The skeleton key of consciousness studies?](#)
- Steve Mashalidis, [Consciousness and education: A process perspective](#)
- Greg Nixon, [The assault of the hard sciences of consciousness on experiential education](#)
- Knut Nordby, [Vision in a complete achromat: A personal account](#)
- Gary Schouborg, [A methodology for the science of consciousness](#)
- John Searle, [Consciousness](#)
- Maxine Sheets-Johnstone, [Consciousness: A natural history](#)
- Keith Sutherland, [The mirror of consciousness](#)
- Mario Vaneechoutte, [Experience, awareness, and consciousness: Suggestions for definitions as offered by an evolutionary approach](#)
- Ken Wilber, [An integral theory of consciousness](#)

- Art Winfree, [Is it impossible to "measure" conscious feeling?](#)
-

Go to:

- [Online papers on consciousness](#)
- [David Chalmers' home page](#)

VISION IN A COMPLETE ACHROMAT: A PERSONAL ACCOUNT

By: Knut Nordby

INTRODUCTION

Being myself a vision scientist who is also a complete achromat, I have, upon demand from friends and collaborators, taken on the task of trying to convey some of my visual experiences and to explain how I cope with my visual handicap. I rely on recollections, both my own and those of my family, and try to separate information that can be documented or is supported by other sources from the mere anecdotal. I also draw on information from interviews with other achromats, but only to supplement or comment on my personal experiences.

My recollections may have become distorted over time, they may not be precise and may also be biased - and they may not, in all cases, apply to other achromats. It is my hope, though, that my account will give both vision scientists and the general reader a glimpse, not only into the manifestations of achromatopsia and rod-vision - topics usually not covered by the learned papers - but also into the practical problems and obstacles encountered by a completely colour-blind person.

A SHORT BIOGRAPHY

Infancy.

I was born on November the 11th, 1942, in Oslo, Norway, as the first child of Kjell Nordby (28.11. 1914 - 30.3.1987) and Mary Camilla Nordby, nee Bredesen, (born 21.12.1914). The pregnancy was normal and the birth went well. My birth-weight was 3845 g and my length 510 mm. There is some early written and photographic documentation showing that, on the whole, I was a healthy baby who seems to have developed normally and, except for a stenosis of the pylorus which was quickly cured, I gave my parents little cause for alarm during the first six months of my life.

Both my parents were born with normal vision and, indeed, excellent colour-vision. I tested both with the Ishihara and the TMsterberg pseudo-isochromatic plate tests, and I also tested my father with the Farnsworth 100 Hue test. They both performed normally. I have one sister (born 8.10.1943) and one brother (born 5.10.1945) who are both, like myself, typical, complete achromats, confirmed by the anomaloscope.

There is no evidence of consanguinity between my parents, although the possibility cannot entirely be ruled out. My paternal and my maternal grandmother both came from the same small coastal town in

southern Norway, but nothing is known with certainty about any relationship between the families. If there is any relationship, it must be at least four generations removed (see Fig. 1).

The prevailing view, then and now, is that typical, total achromatopsia with reduced visual acuity is a recessive, autosomal (i.e. not linked to the pair of X-Y- or sex-chromosomes) inherited condition. The statistical probability, according to Mendelian laws of inheritance, of all three siblings inheriting and displaying a recessive trait from parents who do not exhibit the phenotype is exceedingly small. On the average, and assuming that only one gene is responsible, only one child out of four from heterozygous parents will be homozygous for and display a recessively inherited trait.

No documented case of achromatopsia is known to have existed in either branch of my family. There are, however, rumours of an aunt of my maternal grandmother who, reportedly, had "very weak eyes" and "rarely ventured out-of-doors in full daylight. Despite the suggestive symptoms, such anecdotal evidence cannot be accepted to infer a bona fide case of achromatopsia, since no ophthalmological report is known (it is most improbable that any medical assessment of her eyes was ever performed), and her symptoms can be attributed to several other causes (e.g. retinitis pigmentosa from diabetes).

According to my parents, I could control my eyes and direct my gaze even at the age of three weeks. Photographs taken of me only a few weeks after birth show me with fully open eyes and not a trace of squinting, or partial closing of the eyes, in bright light (see Fig. 2). Nothing unusual was recorded at the time about my visual behaviour (i.e. light aversion or "photophobia") or eye-movements (i.e. nystagmus).

Due to enemy actions in Norway, my parents fled with me to Sweden in May 1943, when I was only 6 months old. One to two months later my parents began noticing that I had developed some strange symptoms. My eyes had started to quiver from side to side (i.e. horizontal, pendular nystagmus) and my eye-movements were irregular. I had begun to blink continuously and partially close my eyes, or squint (looking through the narrow slits between the eyelids), in bright light, and I habitually avoided bright light, something I had not done before. Earlier, according to my mother, I had sometimes even looked straight into the sun with no signs of distress, and she often had to turn me around, being worried that I would damage my eyes. This special behaviour has also been reported to me spontaneously by the mothers of several other achromats. One mother, in fact, believed that this habit of gazing into the sun had harmed her son's eyes and was the actual cause of his achromatopsia and low visual acuity.

After seeing several general practitioners and some ophthalmologists, who could offer no explanation for the symptoms, my parents finally consulted a prestigious professor in ophthalmology (no-one in the family can now recall his name). Hereditary achromatopsia totalis [i.e. total colour-blindness] with concomitant horizontal pendular nystagmus, light aversion, and reduced visual acuity was diagnosed. He also found that I was hypermetropic with a slight astigmatism. At an age of only nine months I thus received my very first pair of spectacles. As far as can now be established these first lenses had a strength of about +3 diopters and had a 1 diopter vertical cylinder.

My parents were told that I was severely visually handicapped, that I was totally colour-blind and that my

visual acuity was so low that I would never be able to read and write. They were also told that I would have to attend institutions for the blind and learn to read Braille, and, at best, could be trained for one of the traditional vocations for the blind (e.g. piano tuner, Telephone switch-board operator, etc.). As it happens, the authority was disproved in some of his predictions by the course of events, but I am getting ahead of myself

After hostilities ended my parents decided to stay in Sweden. Because of my father's work we moved many times; living in Stockholm, in Gothenburg, in the small town T"reboda and again in Stockholm; until the summer of 1954, when his work took us to Caracas, Venezuela. There we stayed four years, until the summer of 1958, after which we returned to Norway and settled in Oslo.

Early childhood.

My first clear memories seem all to be connected with nights and evenings, or they occur indoors in subdued lighting. As far back as I can remember, I have always avoided bright light and direct sunlight as much as possible. Photographs taken of me, and my siblings, during our childhood normally show us with nearly shut eyes, usually looking away from the sun, except when photographers demanded that we look towards the sun for the pictures (see Fig. 3). As a child I preferred playing indoors with the curtains drawn, in cellars, attics and barns or outdoors when it was overcast, in the evenings, or at night.

This, of course, was quite the opposite of what wellmeaning parents and grown-ups considered to be right for children. My whole childhood was, in fact, a continuous struggle against the prevailing views about what, is proper for children, i.e. being out in the sunshine as much as possible, not playing outdoors after it is dark, not drawing the curtains during the day, not playing in cellars and other dark places, and so on.

It soon became apparent to my parents that the learned ophthalmologist was correct insofar as I was completely colour-blind. My father told me that he had noticed that when sorting my coloured playing blocks in day-light I usually put the red and the black blocks together in one pile, the green and the blue ones in another pile, the yellow were usually, but not always, put in a pile by themselves, and so were the white. Indoors in artificial lighting, on the other hand, I sometimes separated the red and the black in different piles, but frequently confused the green and the blue blocks; the yellow and the white blocks, though, I still put in separate piles.

When using colouring pencils or crayons, I am told, I always confused the colours, breaking all the conventions and "rules" about what were the "correct" colours to use: I would happily colour the sky light green, yellow or pink; the grass and leaves orange or dark blue; the sun white or light blue, and so on. I was always corrected in my choice of colours by those who knew better, and, eventually, I gave up painting and colouring my drawings. Unfortunately no coloured drawings from my early childhood seem to have survived.

When I was five we were living in Gothenburg, and from then on my recollections are much clearer. At this time my circle of activity expanded beyond the limits of the house and the garden and into the near

parts of the town. I then developed a system for finding my way back, which I still use. The city-block was the basic unit in my system. I always kept myself orientated where I was on the block. When going further away the system consisted of counting the streets I had crossed, keeping count of right and left turns, counting doors or shops along a block, thus forming a mental, topological map of my path. When returning I reversed direction, retracing my path, making turns in the opposite direction, and counting backwards the streets I had crossed.

I also learned to make use of various prominent landmarks, e.g. parks, squares, churches, towers, underpasses, bridges, etc., as important checkpoints along the route. Even today, I find it much easier to orientate myself in cities with rivers, canals, tramways, overground railway lines and other conspicuous and easily identifiable "boundaries" that divide the city into smaller parts. I cannot remember ever having lost my way returning from a place I had first located myself. On the other hand, it can be more difficult when going with others to keep track of the route, especially when travelling by car.

When five, I nearly fooled my parents into believing that I could read. My parents often read stories and fairy tales to us children, and I was very much intrigued by this activity. I had a children's picture book with a few lines of narrative on each page, all of which I had learned by heart. Sitting with the book at normal reading distance, where, of course, I could not discern the individual letters I "read" the text out loud and clear, following the lines with my eyes, turning the pages at the right points. Eventually, I was found out when, by mistake, I turned two pages at a time and continued to read on before realizing my error. This desire to master the art of reading never left me.

Schoolyears.

When the time came for me to start school we lived in the small village of Töreboda, about halfway between Stockholm and Gothenburg. In Sweden it is compulsory to begin school in the year that children reach the age of seven. The nearest school for the blind and partially sighted was in Stockholm. My parents did not want to send me away to a boarding school at such a young age, and, in defiance of the professor's earlier advice, it was decided that I should try to attend an ordinary school and see if I could manage. Thus, in the autumn of 1948 I started out in the first form of the municipal, primary school in Töreboda.

At this time it became very obvious to me that my vision was different from that of other children. They could see things that I could not see; such as recognizing each other at a distance, spotting ripe berries on bushes and trees, reading cars' licence plates at a distance, etc. They could also take part in activities and sports, especially ball-games, that I could not. Hitting a ball with a bat, or catching a ball thrown towards me, was next to impossible for me, except under the most optimal light-conditions, such as in twilight; When teams were set up, I was always the last one to be taken on.

I must have been a difficult pupil, inquisitive about all that was going on, talkative, always avoiding doing things when my vision would prevent me from performing and thus give me away. I did learn the letters of the alphabet, but I did not really learn to read properly in the first and second form.

The school-books used had ordinary size print, and no-one had thought about getting me books with extra large print, if, indeed, such books existed then. Nor had anyone thought about providing me with a reading-glass. The lenses of my glasses, which by now had been increased in strength to about +6 diopters, did not have enough magnification power for me to read the small print easily.

Since I could actually not discern the individual letters even in ordinary book-print, I again resorted to my old reading hoax; bringing it to new and unprecedented heights. I had developed a very keen memory, and it was usually enough if a class-mate or someone in the household read my home-assignment aloud to me once or twice, in order for me to remember and reproduce it, and to perform a rather convincing reading behaviour in class.

An important discovery that I made during my first school-year is worth noting. As an aid for teaching the letters of the alphabet, the teacher placed large cards, each holding a printed letter, in a row over the blackboard as the letters were introduced. To differentiate between the two categories of letters they had different colours; the vowels were red, while the consonants were black. I could not see any difference between them and could not understand what the teacher meant, until early one morning late in the autumn when the room-lights had been turned on, and, unexpectedly, I saw that some of the letters, i.e. the A E I O U Y Å Ä Ö, were now suddenly a darkish grey, while the others were still solid black. This experience taught me that colours may look different under different light-sources, and that the same colour can be matched to different grey-tones in different kinds of illumination. I have since often used this phenomenon of spectral differentiation as an aid to separate colours by their different grey-tones under various light-sources.

A constantly recurring harassment throughout my childhood, and later on too, was having to name colours on scarves, ties, plaid skirts, tartans and all kinds of multi-coloured pieces of clothing, for people who found my inability to do so rather amusing and quite entertaining. As a small child I could not easily escape these situations. As a pure defence measure, I always memorized the colours of my own clothes and of other things around me, and eventually I learned some of the "rules" for "correct" use of colours and the most probable colours of various things: As an example, I learned that glass that was very dark to me usually was a dark cobalt-blue, glass that looked a bit lighter was usually bottlegreen, and so on. In this way, I could fool some people into believing that I had colour-vision and stop them from pestering me. A friend of the family, though, actually believed that my inability to name colours came from my parents not having taken the trouble to teach me the colour-names. She often tried, but in vain, to train me in naming colours, but had to admit in the end that I could not, in fact, differentiate between the different colour hues.

When I was eight, my three year younger brother was given a small bicycle. This was quite irresistible to me and I soon learned to ride it. At first I only dared to go around the block, keeping to the left-hand side of the road and only making left turns (Sweden had left-hand drive at this time), but as I gained confidence my activity area expanded. Traffic in post-war Sweden was very light, in our village we scarcely saw more than a couple of cars a day, much Transport being carried out by horse and cart, and the conditions of the roads in the area did not allow speeding: Riding the bicycle in traffic did not,

therefore, constitute too grave a danger to me, despite my visual handicap.

I often "borrowed" the bicycle, getting up early in the morning, and set out on long journeys visiting out-laying farms, clay-pits, brick-works, factories, railway stations and all kinds of interesting places. When not chased away at once, I would often stay for hours. My curiosity was unbounded, and I learned many wonderful things; how bricks are made, the working of farm machinery, railway operations, etc. When visiting the local railway station the engine driver found it was safest for me (and best for his peace of mind) to take me with hi'm in the cab of the locomotive, rather than have me nosing around the tracks - to which I did not object at all. This inquisitive self-education may have been the origins of my later interest in research.

In the summer of 1950 we moved to Stockholm and I started the third form in a large, municipal school in one of the newly built suburbs south of the city. This did not work out very well. With large classes, the teachers had little time to give me the special tuition and extra training I needed, and after only a month or two it was finally decided that I, and my sister, who had started school the previous year, should attend the state school for the blind and partially sighted at Tomtebodavägen, just north of Stockholm (my brother was not due to start school until the following year). Because the trip each day between the home and the school took too long, my sister and I were forced to live in, seeing our family only in the week-ends and during vacations.

At this school my sister and I were, in all practical matters, treated as blind. Although the staff must have known that I actually could see quite well in some situations, a regime was set up to teach me to write and read Braille. Reading with the tips of my index-fingers I did not find easy and I quickly developed the knack of reading Braille by eye, because the raised dimples of the Braille letters cast shadows on the paper, making them much easier to read by sight than by touch. For this I was punished; it was considered to be cheating, and for a week I was confined to my bed after classes, being denied to see the other pupils. And, for the next couple of months I had to wear a heavy, lined, black velvet mask covering my eyes in class when reading - to keep me from peeking. Eventually my Braille reading proficiency increased to such a level that the mask could be dispensed with.

Although I had in many ways become a leader at the school for the blind, having a definite visual advantage over the genuinely blind and the gravely partially sighted pupils, and thus gained much confidence in dealing with other children, I was quite unhappy there and strongly resented being treated as blind. So one day I ran away. Completely on my own I crossed Stockholm, from the north to the far south of the city where we lived. I reached my home by midnight, after more than ten hours of walking, retracing the exact route my father used to take us home by car on the week-ends. This caused a great scandal: Unbeknownst to me the bishop was to hold his visitation of the school on that very day, but the staff had to search for me and the police had to be called in to help them, thus ruining the whole event. This in no way improved my popularity with the staff; although, as was later somewhat reluctantly admitted, it was considered to be no mean feat, even for a child with normal vision, to find his way across the entire city of Stockholm.

Thus, after only two years at the school for the blind and visually disabled, it was decided that, as a trial, I

should again try to attend an ordinary school for the sighted. I was transferred to a newly opened municipal, primary school very close to where we lived, although it meant being moved down one form to compensate for all I had lost during my two years at the school for the blind.

I have a strong suspicion that at the school for the blind and partially sighted they actually just wanted to get rid of me, since I had always created much turbulence and was considered to be rather detrimental for morale among the more seriously handicapped pupils. My sister and my brother (who had entered the school for the blind the previous year) both continued to attend the school, though, staying for two more years until we left Sweden in 1954.

Seen in retrospect, the curriculum at the school for the blind may appear to have been extremely oldfashioned, but it should be borne in mind that at the time this school was considered to be one of the world's leading institutions for the blind and partially sighted, and it was run according to the most modern pedagogical principles. Today, 35 years later, we can only be happy that some of these principles now seem to have been abandoned.

During the summer holidays of 1952, before joining my new school, a problem troubled me very much; although I told no-one about it: While at the school for the blind I had forgotten practically all of what little I had learnt before of reading ordinary print, and I did not cherish the prospect of facing my new fourth form class not being able to read and write.

My father, being at that time sales director with a large company, kept a stock of handouts and gifts for customers, which among other things contained lighters, pocket knives, and -pocket magnifiers. One day I borrowed a magnifier and used it to look at pictures in a comic-book. I was frustrated at not remembering how to read the text in the speaking bubbles, since I could now clearly discern the individual letters with the help of the magnifier. I then remembered that my parents had been given a sample-sheet with the Braille alphabet embossed on it together with the ordinary printed alphabet. Using the sheet as a "cipher-key", I managed to "decipher" and read the text in the speaking bubbles.

This incident was truly a turning-point for me and, more than anything else, opened a whole new world to me, who had always so much wanted to be able to read. In just a few weeks' time I secretly taught myself to read properly and my reading skills very quickly improved. I became a voracious reader who devoured all I could lay my hands on, buying magazines and borrowing books in libraries on practically any imaginable topic.

In my new school I had the unforeseen fortune of having a teacher who had an understanding for my handicap and who really tried to help me overcome my problems in class. He was in many ways a very progressive teacher, ahead of his time. He did not adhere to the strict regime practiced in the schools in those days. As an example he let me move my desk up close to the blackboard so that I could actually see what was written, breaking the regular geometric order of the desks in the classroom, then so much cherished by the other teachers, and he let me move around in the classroom during lessons so that I could inspect at close quarters things that were demonstrated; something completely unthinkable with his

colleagues who put their pride in having their pupils sitting quietly at their places. And he made a habit of always telling us what he wrote on the blackboard, which was of great help to me.

Even today, when attending lectures and talks, I always try to sit close to the screen when slides are shown, and I find it easier to follow the presentation when the speaker also tells the audience what is written on the transparencies. When lecturing or giving talks myself, I always do this, showing slides with large print and not crowding too much information into each transparency. In fact, I often treat my audience as visually disabled: If I can see well from a front row seat what I am presenting I know that my audience also will see it.

Because of this outstanding teacher I soon managed to catch up with my peers, regaining much that was lost in the years at the school for the blind. It was also very good for my self-esteem to be able, at least in some fields, to perform like other children of my age, and it also encouraged my parents to try sending my siblings to an ordinary school.

As it happened this was to take place in Caracas, Venezuela. For four years my siblings and I attended a private school run there by a Scotsman. It was a very small school, and we had no great problems in learning English and otherwise adapting to his somewhat unorthodox style of teaching.

After returning to Norway in 1958 we consulted the Department of Ophthalmology at the National Hospital in Oslo. Here Dr. Egill Hansen confirmed the earlier diagnosis of complete achromatopsia with reduced visual acuity, light aversion and nystagmus in all three of us using the anomaloscope and the standard battery of tests - and we were entered into the Archive of the Blind. In Norway, like in many other countries, people with visual acuities of 6/60 (1/10) or less are legally considered as blind. This entitled us to special education and to pensions for the disabled, which we have never made use of.

Instead we started to attend grammar school (gymnasium in Norway) where we each, in turn, sat for the General Certificate of Education, Advanced Level (which is required for going to the university). At exams we were given an extra hour for finishing our papers, but little else was done by the school authorities to facilitate us attending a school- for sighted pupils. By now, however, we had all become rather expert in coping with the practical problems of attending a school for the sighted. Upon getting our exams we all, in turn, went on to university.

At this time I started my musical education, taking lessons on the clarinet, the cello, and in theory of music at the Oslo Conservatory of Music. I did not have any special plans of becoming a musician, but acquired a professional education along with the Gymnasium. Reading printed music proved to be a real obstacle for me, especially *prima vista*, i.e. sight reading. Today I can easily make enlargements of sheet-music on an office copier, but in the early 60's this was not available to me. For soloists this may not be a great problem, since they often have to learn their parts by heart anyway, but in orchestra play it can be a great handicap.

At university I didn't do too bad. From early in my studies I was employed by the University as part-time

lecturer, which meant that I could have my own office. There I had full control of the light level, drawing the curtains during the day if necessary to shut out sun-light, and I could thus work under the conditions that suited me best. When attending lectures and seminars I made a habit of arriving early, not out of courtesy to the lecturer, but to be able to get a good seat, sitting, if possible, with my back to the windows and close to the blackboard or projection screen.

After first having a go at law, which I gave up after one year, I eventually ended up studying psychology at the Institute of Psychology at the University of Oslo. In addition I studied philosophy of science, physiology and music science. My interests soon turned to sensory psychology and perception. As I wanted to work independently, I took up auditory research, partly out of interest for audition, but also because I believed that my visual handicap would not be a big obstacle to me in this field of research.

At this institute also worked Professor Ivar Lie, Svein Magnussen, Bjorn Stabell and Ulf Stabell, all very accomplished vision scientists. They realized that, being trained in psychophysics, I would be a unique subject in research on rod vision, and they tried to persuade me to leave audition for vision. I first wanted to get my Ph.D. (Mag. art. in Norway), and after my dissertation (on auditory localization and dichotic time/intensity-trading) I switched to vision research.

I have been in research ever since; first as assistant professor (stipendiat in Norway) at the Institute of Psychology, University of Oslo, working in vision research and lecturing in perception and neuroscience, and today as senior research scientist with the Norwegian Telecommunications Administration Research Department, where I am currently doing research and development work on human factors in telecommunications, on telecommunication equipment and services for disabled people and in telemedicine.

My vision research activities soon brought me into contact with vision researchers in laboratories outside Norway, some of which I have visited many times and where I have made many good friends. Beside work in Norway with Svein Magnussen and Bjorn and Ulf Stabell, I have also visited and collaborated with Dr. Robert E Hess at the Physiological Laboratory, University of Cambridge, England, with Dr. Lindsay T. Sharpe, in Dr. Lothar Spillmann's laboratory at the Minikum der Albert-Ludwigs Universität, Freiburg, the Federal Republic of Germany, and with Dr. L. Henk van der Tweel, at the Laboratorium voor Medische Fysica, Universiteit van Amsterdam, The Netherlands - collaborations that have continued after I joined the Telecommunications Administration. The results of these experiments are dealt with in other chapters of this book (for a full list of these works see the Bibliography below).

My siblings also entered university; my sister studied Latin, English and music science, and my brother qualified as a psychologist. My sister now works at the library of the Norwegian Association for the Blind in Oslo, translating and producing books in Braille; unlike me she has not forgotten her four-year training at the school for the blind. My brother is employed as a high-ranking civil servant with the Norwegian Directorate for Labour, working mainly in organizational development.

In 1965 I met Nina Marie Loberg (born 10.7.1944), who studied psychology, and we were married the

following year. There are no known achromats in her family and there are no relations between our families. We have two sons, Cato (born 1.3.1967) and Alexander (born 19.1.1969). Both Nina and the boys have perfectly normal colour-vision as tested with the Farnsworth 100 Hue Test. My brother is also married and has four children, two sons and two daughters; all with normal colour-vision. My sister has not married and has no children.

I have never, nor have my siblings, received any qualified vocational counselling or help in planning my education and my career. What advice I have had has usually been unfounded and ill-conceived. The suggestions have included everything; from simple clerical and assembly-line work to the traditional work for the blind, usually involving simple, manual, low-paying jobs requiring little or no education, and showing complete disregard for what actual tasks the work demanded. Most so-called 'simple' jobs, e.g. filling in forms, sorting mail, filing and retrieving files, assembling components, cleaning, tending machines etc., usually require good visual acuity and colour-vision. No one ever suggested theoretical or academic work requiring higher education. Luckily my parents had always wanted their children to have an academic education, and circumstances were highly favourable for this in Norway in the 60's and 70's.

In my research work I have had few problems related to my visual handicap. I work in areas where, to a large extent, I can plan my own work and choose my own tasks. In addition to the research activities my work also includes administrative duties, giving interviews, going to meetings and giving talks, both in Norway and abroad, and I also lecture regularly at the University of Oslo. These activities pose few practical problems for an achromat. Typing used to be quite a bother, since I had to lean over the typewriter with my magnifier to be able to read what I had written, but now I use a PC and a word-processing program for all my typing, and this is a very useful tool to me. To overcome my low visual acuity I have to get close (150-200 mm) to the screen, and I have a large (19") screen installed. With modern printers I can have my documents printed in large type-face, making them easier to read.

ROD VISION

Trying to explain to someone with normal, or nearly normal, colour-vision what it is like to be totally colour-blind, is probably a bit like trying to describe to a normally hearing person what it is like to be completely tone-deaf, i.e. not possessing the ability to perceive tonal pitch and music. My task, though, is probably a bit simpler than the case of the tone-deaf, since practically everyone has had experiences of achromatic (i.e. colourless, or black & white) or monochrome pictures and renderings, and certainly must have witnessed the gradual disappearance of colours when darkness sets in.

A first approximation, then, in explaining what my colour-less world is like, is to compare it to the visual experiences people with normal colour-vision have when viewing a black & white film in a cinema or when looking at good black & white photographic prints (good here meaning sharply focused, high contrast with a long grey-scale, as in crisp, high quality, glossy, technical prints).

This, however, is only part of the story because I have so far only dealt with the achromatic aspect of my

perception. To get a fuller understanding of my visual world one must, in addition to my colour-blindness, also take into account my light aversion (i.e. hyper-sensitivity to light) and my reduced visual acuity. In the following I will deal in turn with each of these aspects of typical, total achromatopsia.

Total colour-blindness

As mentioned above, I only see the world in shades that colour-normals describe as black, white and grey. My subjective spectral sensitivity is not unlike that of orthochromatic black & white film'. I experience the colour called red as a very dark grey, nearly black, even in very bright light. On a grey-scale the blue and green colours I see as mid-greys, somewhat darker greys if they are saturated, somewhat lighter greys when unsaturated, like pastel colours. Yellow is usually a rather light grey to me, but is usually not confused with white. Brown usually appears as a dark grey and so does a very saturated orange. When asking other total achromats to do such matches I get practically the same judgements from them. In the literature I have also found several descriptions of the colour-determinations made by other typical, total achromats, which all very closely match my own judgements (see e.g. Bjerrum, 1904; Larsen, 1918).

Although I have acquired a thorough theoretical knowledge of the physics of colours and the physiology of the colour receptor mechanisms, nothing of this can help me to understand the true nature of colours. From the history of art I have also learned about the meanings often attributed to colours and how colours have been used at different times, but this too does not give me an understanding of the essential character or quality of colours.

Coloured and B & W pictures are usually indistinguishable to me. But sometimes I can, often quite easily, tell coloured and noncoloured pictures apart. Coloured pictures may look less crisp, or slightly less in focus, and often have less contrast than comparable monochrome pictures. Under some conditions I am able to tell a polychrome picture from a monochrome one by noting the different surface textures of the coloured inks used to print the picture. I do this by tilting it so that the light reflects differently in the different inks: Coloured inks of various hues show up as duller or glossier patches. I am sometimes able to see the numbers on the Ishihara and TMsterberg pseudo-isochromatic plate tests by this technique. Given the time I might be able to fool an unwary tester. A few other achromats have also shown me this trick, not because they wanted to deceive me, but because they thought that they had a vestige of colour-vision. However, if I cover the test-plate with a glossy, transparent foil they are unable to do this trick: Varnishing or covering pictures with high-gloss films destroys the effect.

I have never experienced anything "dirty", "impure", "stained" or "washed out" about colours, as reported by the artist Jonathan I. who completely lost his colour-vision after a cerebral concussion (Sacks & Wasserman, 1987). When, occasionally, colour-pictures do look less crisp to me than B & W pictures, it is usually because of bad printing, lower contrast or faded inks.

Once when Bjorn and Ulf Stabell were measuring my spectral equal-brightness function (using their Wright colorimeter) I felt that I could detect slight differences between colour hues by distance cues,

which could be due to chromatic aberration in the optics. I informed the Stabell brothers about my suspicion and an ad hoc experiment was quickly set up to test my conjecture. I was to try any ploy that I could think of to guess the hue of the test field, and they were to counter my ploy, making mock changes or unexpected changes of hue. When the day was over I had not been able to guess any colour hue beyond the pure chance level.

When I know the colour of an object I may often refer to its colour name when describing it or referring to it to other people. Since colour names have meanings for most people, communication is made easier. This, in turn, leads to other people using colour names when addressing me, even those who know of my colour-blindness. Sometimes this can be of help to me, when e.g. they refer to a red book among other light-coloured books, but re'ferring to a red book among black and dark-coloured books is of little help to me.

Colours do not help me to distinguish objects from their backgrounds. Since the grey-tone that I can match to a particular colour changes with changing illumination, objects that are partly in bright light and partly in shadow, such as flowers in the mottled sunlight coming through the branches of a tree, can be very hard to see. The same applies to objects under glass, behind windows or under water, since shiny surfaces give off reflections and glare which makes it very difficult to see what is behind them. It is usually impossible for me to recognize people in cars or see if they are waving at me, sometimes even at close quarters, because of the destructive reflections from the windows. NMere the colour-normal can use the continuity of the colour-hues of an object to perceive the figure through disturbing reflections, I see the reflections and the objects behind them translated into different shades of grey, and only when the contrast between the object and its background is very high, or the movements are very distinct, can I see what is behind a strongly reflecting surface as an object. Wearing polarizing clip-on sun-glasses improves things to some extent, and I use them much, although I am sometime's distracted by seeing the mottled stress-patterns in glass that are normally invisible.

When I was about 14 I discovered, quite by accident, the use of coloured filters for analyzing and identifying different colour hues. We had a large salad bowl made of transparent, red-coloured plastic. One day, holding the bowl before my eyes, I noticed that the pattern of the table-cloth turned nearly white when seen through the bowl. By moving the bowl to and fro, I could make the red-coloured pattern change from light to dark. I was intrigued by my discovery and tried out several other transparent filters of various colours, thinking that I would be able to solve the problem of my colour-blindness by using a few differently coloured pieces of plastic to analyze and thus be able to name any colour. From a purely practical point of view I contemplated the great advantage of being able to determine for myself the name of a colour without having to ask other people. I even envisaged a solution of narrow strips of filters mounted at the upper edge behind the lenses of my glasses, thus dispensing with a set of hand-held filters, and making their use more discreet.

Nothing came out of these ideas of mine as I did not have the resources at my disposal for ear"rying them through. I have since learned that such filters for analyzing colours are actually made for "normal" (i.e. anomalous trichromat and dichromat) colour- blind electronic engineers to aid them in reading the colour-coded rings denoting the values of resistors. I have tried, but never succeeded in, obtaining a set of such

filters, although the spectral transmission bands of these filters would probably not suit the spectral sensitivity of my rods very well.

Traffic-lights can sometimes pose a problem to me. I have long since learned that red is the top light, amber is the middle light and green is at the bottom. At night, in the evening or in the day if the trafficlghts are in a shadow, I have no problem in detecting which aspect is shown: By noting its position I can be a law-abiding cyclist or pedestrian. In full sun-light, on the other hand, determining which aspect is shown can be quite impossible; if the traffic-lights are back-lit I am dazzled by the sun and cannot see the weak lights; if the sun is behind me, the reflections in the lenses of the light signals can be so strong that it is impossible to determine which one is showing. In such situations I have to watch the traffic, following the other pedestrians when they cross the street. This can lead to dangerous situations when encountering unsuspect'ed cars if I follow someone crossing a street against a red light, believing that the crossing is safe.

I often visit art galleries and I am very interested in the visual arts. Looking at paintings I can appreciate form, composition and technique, although of course, I cannot appreciate the coloristic aspects. Monochrome prints and graphic art I can normally enjoy in the same way as colour-normal people and the same goes for sculpture and architecture. When lecturing on vision and perception I frequently show transparencies with works of art to demonstrate my points.

When painting house or redecorating I must always make sure that I use the correct colours by carefully reading the labels on the paint tins. My wife normally selects the colours for the house. If the contrast between the new and old colour in terms of grey-tones is not too small, I have no problem painting the house. If I repaint with the same colour, I sometimes have a problem telling the freshly painted areas from the unpainted ones, especially after a break, when the fresh paint is not so 'wet' anymore. If I repaint with a different colour that is very close to the old one in terms of grey-contrast, I have difficulties in seeing if the new coat covers the old colour properly. Sometimes the only way to resolve this problem is either to paint very systematically, or to look obliquely at the wall for the shiny 'wet' freshly painted parts.

When buying clothes on my own I will only take advice from a sales-person in whom I have the highest confidence, otherwise I usually ask for "safe" or neutral colours; white shirts, grey trousers, black socks and shoes: I would e.g. never select a tie on my own. For the important colour-choices I have to rely completely on my wife or on close friends who I trust and who know my taste and preferences. To avoid embarrassment I often have to mark my socks in some way so that I will not mix up differently coloured pairs, it can be impossible for me to tell light blue, beige and light grey, or dark blue and black socks apart.

Picking berries has always been a big problem for me. I usually have to grope around among the leaves with my fingers, feeling for the berries by their shape, except in the shade or in the evenings when light levels are low. Then I can usually see the berries; the red ones as small 'black' spheres among the 'grey' leaves. The only berries I can easily see in bright sun-light are the white 'snow-ball' berries of the Guelder-rose bush.

Indoors most flowers are easy for me to discern, out-of-doors, though, I usually only recognize the white and the yellow ones: The oxeye-daisies, not to speak of the dandelions in my lawn, are most conspicuous to me; red roses on bushes I see best in the twilight when I can separate the "black" flowers from the "grey" leaves.

Colours are frequently used to code or to highlight information. For me this usually makes matters worse because good colour contrast very often does not transform into good grey-contrast. Sometimes the contrast is so low that the information is next to lost. Black print on red price labels, yellow print on a light blue background, dark green on a bright red background are all extremely difficult for me to read. A teaching aid that has always failed with me is the use of colours in e.g. arithmetic; adding the red and the green objects, subtracting the blue ones from the yellow ones. Colourful Venn diagrams for teaching basic set theory only makes the task more difficult for me.

Hypersensitivity to light

As far as can be determined, the retinae of my eyes do not contain any cone-receptors at all, only rod-receptors; or cones are present but in such reduced numbers that they do not in any measurable amount contribute to the visual process (see Sharpe and Nordby, THE PHOTO-RECEPTORS IN THE ACHROMAT, chapter X). Since the rods are much more sensitive to light but also saturate at lower light intensities than the cones, my visual system is well adapted for vision only under low-lighting conditions. In fact, my vision will not function at all in bright light (e.g. out-of-doors in full day-light), if I do not adopt specialized visual strategies and behaviour.

I am easily dazzled and, in effect, blinded if I expose my eyes to bright light. If I open my eyes fully for more than one or two seconds under such conditions (about 1000 scotopic trolands and up), the scene I am gazing at is quickly washed out, turned into a bright haze and all structured vision is lost. It can be very distressing for me, and sometimes even painful, to perform demanding visual tasks in very bright light.

This hypersensitivity to light, or light aversion, is usually referred to as photophobia, but has nothing to do with the irrational psychodynamic "phobias". In fact, I really like being out in the warm sun, provided I don't have to perform exacting visual tasks. I do not like to read or write in the sun; digging the garden or mowing the lawn, on the other hand, is no problem for me even in bright sun-light.

My main problem, then, is to restrict the intensity of the light entering my eyes. This, as I will show, can be achieved in several ways and I tend to use them all, alone or in combination, according to what the situation demands. In common with nearly all the other achromats I have interviewed, I have developed special visual strategies for restricting the amount of light entering my eyes.

The most obvious strategy is, of course, simply to avoid direct strong light. Staying indoors or in the shade is one way of achieving this if there is no special reason for being in intense sunshine (such as on a

bright beach in the summer or in bright sun-lit snow in the winter). Indoors, whenever possible, I try to place myself with my back towards bright windows and strong light sources and avoid having direct sunlight falling on my work-place.

Shading my eyes from direct, intense light with my hand or a visor may be necessary. Ordinarily I wear sun-glasses, in casu, the lenses of my glasses are photo-chromatic (i.e. tinted lenses which darken in bright light). Out-of-doors I often wear an extra clip-on polarizing filter for cutting down on bright light and for dealing with visually destructive glare and reflections from shiny surfaces. The best sun filters I have tried, though, are the special coloured glasses which are made for the retinitis pigmentosa patients. They have a spectral cut-off at 550 nm, passing only the long wavelengths and they give a very pleasing light-attenuation, but they are socially less attractive because of their red appearance.

Actually, it is not only the brightness of what I am looking at that is most bothersome, but the brightness of the total visual field; the larger the part of the field that is illuminated, the more bothered I am. Blinkers or side-shields on the spectacle frame may help in shutting out unwanted light, but they also prevent motion detection in the peripheral visual field, which is important for moving safely about, and I don't use them. If I must see small detail in bright light, e.g. read printed information or look at a map in the bright sun, I turn away from the light, putting the material in my shadow.

The most typical visual strategies, though, that I resort to consists in squinting, i.e. partially closing the eyes and looking through the narrow slits formed by the eyelids, and in frequently blinking the eyes. This habit has been reported by nearly every author and seems to be universally resorted to by typical, complete achromats (see e.g. Bjerrum, 1904; Larsen, 1918; Krill, 1977). If light levels rise and my retinal illuminance approaches 1000 scotopic trolands, my fully constricted pupils cannot further contract and I have to squint to deal with the higher intensities and to avoid saturating the rods.

At higher light intensities even this is not enough, and I also have to start blinking my eyes to shut out excessive light. My blinking is triggered when saturation sets in. The blinking frequency is slow at first, only once every four-five seconds, but the periodicity increases with increasing light intensity, to three to four blinks a second.

At lower light levels, when the blinking rate is slow, the blinks themselves are also rather brief. As light levels go up, the blinking rate also goes up, but the blinks become longer, i.e. the duty-cycle increases. At the very highest light levels where my visual system will function (e.g. new snow in bright sunlight) the blinks are so long that my eyes are, in fact, closed most of the time except for brief opening blinks once every two to three second. Whether blinking my eyes briefly at low light levels or extending the length of the blinks at high intensities, I still experience a visually stable world in which I can orientate myself and move about. It should be mentioned, though, that I am able to suppress the blinking on some occasions. In laboratory investigations I have actually sustained luminances over 500,000 photopic trolands, blinking only occasionally to keep my corneas moist.

This squinting and blinking behaviour is a strain socially. In bright light people immediately notice that

sometl-iing is wrong with my eyes. and show this by their reactions to me. As a child I was often approached by complete strangers who wanted to know what was wrong with my eyes. Wearing dark glasses or clip-on sun filters can alleviate this social burden to some extent.

At higher light levels the peripheral visual field is much more affected than the central part of the field, which results in a partial tunnel vision. I will still detect movement in the far periphery, but have much more difficulty in identifying what is moving and reacting adequately to it. This makes me move in a rather hesitant and stiff way, sometimes bumping into people or things, and to be overcautious when moving about in agitated surroundings or in strange places. As soon I am in the shadow or indoors again I move in a more relaxed and confident manner.

It is very clear to me today that the most debilitating, handicapping and distressing consequence of my achromatopsia has been my hypersensitivity to light and the resulting light aversion, which is also reported by all other authors. This is also the unanimous verdict of all the achromats I have interviewed so far. Both the practical problems of being dazzled, the narrowing of the visual fields, resulting in restricted mobility, and the social problems of light aversion and feeling clumsy in intense light is frequently reported by my informants as being more of a hindrance to them than not being able to experience colour hues or to discern small detail.

Reduced visual acuity

In a retina where only rods function, visual acuity will be drastically reduced, since it is the densely packed foveal cones that are responsible*for the high acuity of central vision in the normal retina. My visual acuity is 6/60 Snellen (i.e. 0. 1 of normal acuity) which means that at a distance of 6 meters I can read those letters on the Snellen chart that people with normal acuity can read at 60 meters. This low acuity is typical for all typical, complete achromats reported in the literature and for those I have interviewed. Measured with an interferometer under optimal conditions my acuity improves to 6/50.

My visual acuity, though, may vary quite a bit depending on the illumination. At higher luminances my acuity very quickly deteriorates, but it may improve slightly at lower light intensities, such as at dusk after the sun has set, but before it gets really dark, or indoors with the curtains drawn during the day, or in the evening with not-too-strong incandescent, tungsten illumination. My best acuity seems to coincide with the illuminance levels at which I do not have to squint or blink my eyes.

I experience a visual world where things appear to me to be well-focused, have sharp and clearly defined boundaries and are not fuzzy or cloudy: I can easily tell the difference between what people with normal vision call a well-focused and a not so well-focused photograph. When giving a talk to the Kenneth Craik laboratory in Cambridge, one of my slides came up slightly out of focus because of a thinner slide-mount. I quickly refocused the slide myself, but Professor Fergus W. Campbell, who attended the talk, commented aloud that to his astonishment I did this as well as any projectionist. Details that are too small for the low resolution of my coarse retinal matrix disappear, blending into the background, but when they are brought closer and become large enough for me to discern, they are just as sharp and well-defined as

other objects.

I am hypermetropic, or far-sighted, and my ophthalmic correction today is +8 diopters. This is rather fortunate since my lenses thus magnify the retinal image by a factor of 0.7; it would have been worse had I been myopic, or near-sighted, thus requiring minus-lenses. With my glasses I can read small print by bringing the text close up to my eyes; about 250 mm away for 6 mm high letters, 150 mm away for 4.5 mm letters, 80 mm away for 3 mm letters.

Letters smaller than 2 mm, such as used in the telephone directories in most countries, I cannot read without additional optical magnification.

I use a small, folding, pocket magnifier for reading newspapers, magazines, books, typed documents, etc., even when the print is large enough to read with my glasses alone. My first pocket-magnifier (see above) had a +9 diopter, 50 mm o lens set in a transparent perspex frame, rivetted in one corner to a piece of leather. The lens could be folded up into the leather cover to protect it, when unfolded the leather cover served as a handle for holding the glass. Today I use a magnifier with a +16 diopter, 30 mm o, biconvex lens, which gives me the necessary magnification for reading large amounts of text without undue visual strain, and makes it possible for me to manage the very small print in Telephone directories etc. When reading I usually hold the glass in my right hand, the book in my left, and I always use my dominant left eye. Together with my spectacles, this magnifier is the most important tool I have, and I always carry it with me.

I have, of course, tried many other different loupes, reading glasses and magnifiers of various designs offered by opticians, but they are usually either too big, too heavy, too obtrusive, not pocketable, too weak or too strong for my needs. The best design is one that I can carry in my shirt breast pocket so that it is always at hand, one that is small and easy to hold, but with sufficient power for reading small print such as the print found in the Telephone directory.

I have considered other types of visual aids, such as video-systems etc., but they all seem to fail on one or several points. Usually they are too large and heavy to carry around, they are complicated in use or they are too powerful, giving a very small field of view which makes it rather hard to read continuous text. It is most important to me that I can always take the aid with me and use it wherever I go.

Reading things at a distance, such as street names, destination boards on buses and trains, time-tables in stations, flight-departure indicators in airports, signs and labels in museums, price labels in shop windows etc., is usually impossible or, at best, very difficult. To use my magnifier to make the letters legible to me I have to get close to the material, but this is not possible when signs are placed high up on walls or poles, behind windows, in glass cases or behind barriers. One special problem is reading the destination boards on approaching buses which only stop on signal. Occasionally, as a last resort, I have had to stop every approaching bus until the right one came along, invoking the drivers' wrath and being showered with abuse for my pains.

It can be difficult for me to find my way in unknown surroundings, a problem when travelling or visiting strange places. I solve this by carrying with me a small, monocular, 8 power, close-focusing telescope, which I can conceal in my hand, and which I use for reading street names, destination signs and other information that I cannot otherwise reach. I also procure information in advance by studying maps and plans, and by inquiring for detailed directions on how to find my way.

Much more serious, though, than the problem of finding my way in strange places, is my disability to identify people only from their facial features at a distance of more than a few metres. I may easily pass people I know well in the street without recognizing them, and those who do not know of my visual handicap may find me aloof or downright rude. Picking out people I know well in a crowd or in a large room, e.g. a restaurant or theater, is very difficult and socially embarrassing. I usually recognize people from their total appearance, i.e. their clothes, their way of moving, prominent or special features; but foremost, from their voices. @en expecting people I know well I can manage to identify them at 10-15 metres distance, and if I know what clothes they will be wearing I may even be able to identify them at distances of 20 metres and more. Often, though, I fail to recognize people I know, with embarrassing consequences. Sometimes even people who know about my visual handicap are offended by my seemingly disinterest and neglect of them. The reason may be that I often, apparently, behave visually rather normally, leading them to believe that my disregard is intentional.

A problem that seemed more serious to me earlier than it does now is that I cannot hold a driver's licence. My brother and I cycle and we have ridden mopeds in Oslo. Driving a car, though, is not possible for us, although my brother actually took some driving lessons. He had to give up, though, after attempting to drive down some stairs in a park. Today I only cycle where I have access to reserved cycle lanes or when traffic is light.

In other fields, however, low visual acuity has not been an insurmountable obstacle. My sister has since her early teens done all kinds of needlework, producing tatting, bobbin-lace and embroideries to very high standards. To keep her hands free, she uses a clip-on, jeweller's loupe with her working- glasses. This gives her a large magnification (10 times), but a very small field of view, and it is not very suitable for reading. She needs help in selecting the colours of the threads and yarns, but once they are selected and labelled she manages on her own.

CLOSING REMARKS

From my account it can be seen that I often avoid showing my visual handicap and sometimes simulate having normal vision. I often experience negative attitudes if I expose my visual disability before people get to know me. The stigma of being partially sighted can be very unpleasant. If I show my handicap early to people who do not know me I will be categorized as a disabled person and will be treated in a patronizing way and often not taken seriously. If, on the other hand, I first expose my visual handicap to people after they get to know me, I get fewer reactions. With most vision scientists and ophthalmologists I feel at ease, they can even be overcautious about offending me, but some ophthalmologists have treated me in the condescending way they sometimes treat their patients.

In spite of this and all the other practical and social problems I encounter, I still feel that I live a very rich and interesting and useful life. It is my ambition that these very personal and rather private accounts given here will prove to be of some use to vision researchers and to people dealing with typical complete achromats.

ACKNOWLEDGMENTS

I wish to thank my colleague and friend Dr. Lindsay Theodore Sharpe for his critical comments on the manuscript. I also want to thank my collaborators and friends Bjorn Stabell and Ulf Stabell for many helpful suggestions.

FIGURE CAPTIONS

The pictures are not implemented, because of their bad quality (copied).

Figure 1. My family pedigree. Solid, black symbols indicate the typical, total achromats, i.e. my siblings and I, light grey symbols indicate non-achromats.

Figure 2. Photograph taken of me 1-2 weeks after birth. I showed no obvious signs of squinting or light aversion at the time.

Figure 3. Photograph of me, my sister and my brother (from left to right) in bright sun-light. We all display squinting and light aversion behaviour typical of the complete achromat.

Figure 4. Spectrum with grey-matches in sun-light (upper) and in incandescent light (lower row). The matches of the grey patches are made at the points indicated on the spectrum. The Spectrum was obtained from the Rochester Institute of Technology, the grey patches are from the KodakLyrev scale.

BIBLIOGRAPHY

The following is a list of all the investigations of my visual system that have been published.

- **Greenlee, M. W., Magnussen, S. & Nordby, K. (1988)**
Spatial vision of the achromat:
Spatial frequency and orientation specific adaptation.
Journal of Physiology (London) **395**, 661-678.
- **Hess, R. E & Nordby, K. (1986a).**
Spatial and temporal limits of vision in the achromat.
Journal of Physiology (London) **371**, 365-385.
- **Hess, R. E & Nordby, K. (1986b).**
Spatial and temporal properties of human rod vision in the achromat.

- Journal of Physiology (London)* **371**, 387-406.
- **Hess, R. E, Nordby, K. & Pointer, J. S. (1987).**
Regional variation of contrast sensitivity across the retina of the achromat: Sensitivity of human rod vision.
Journal of Physiology (London) **388**, 101-119.
 - **Nordby, K. (1988).**
Sosiale mekanismer, stemplingseffekt m.v. ved synshandikap.
Skolepsykologi **239** 1-8, (ISSN 0333-0389).
(Soeial mechanisms, stigma etc. connected with visual handicaps.)
 - **Nordby, K. & Sharpe, L. T. (1988).**
The directional sensitivity of the photoreceptors in the human achromat.
Journal of Physiology (London) **399**, 267-281.
 - **Nordby, K., Stabell, B. & Stabell, U. (1994).**
Dark-adaptation of the human rod system.
Vision Research **24**, 841-849.
 - **Rosness, R., (1981)**
Corticale konsekvenser av stavmonokromasi.
En psykofysisk studie av orienteringsselektivitet.
(Master's Thesis, Institute of Psychology, University of Oslo, 1981)
ISBN 82-569-0541-7, 83 pages, 12 figures.
 - **Sharpe, L. T., Collewijn, H. & Nordb K. (1986).**
Fixation, pursuit and optokinetic nystagmus in a complete achromat.
Clinical Vision Sciences **1**, 39-49.
 - **Sharpe, L. T., Fach, S. & Nordby, K. (1988a).**
Temporal Summation in the achromat. *Vision Research* (in the press).
 - **Sharpe, L. T., Fach, S. & Nordby, K. (1988b).**
Rod increment and flicker thresholds in the normal and achromat.
Perception **17** (in the press).
 - **Sharpe, L. T., Van Den Berg, K., van der Tweel, L. H. & Nordby, K. (1988).**
The pupillary response of the achromat.
Clinical Vision Sciences (in the press).
 - **Sharpe, L. T., van Norren, D. & Nordby, K. (1988).**
Pigment regeneration, visual adaptation and spectral sensitivity in the achromat.
Clinical Vision Sciences **3**, 9-17.
 - **Skottun, B. C., Nordby, K. & Magnussen, S. (1980).**
Rod monochromat sensitivity to sine wave flicker at luminances saturating the rods.
Investigative Ophthalmology & Visual Science **19**, 108-111.
 - **Skottun, B. C., Nordby, K. & Magnussen, S. (1981).**
Photopic and scotopic flicker sensitivity of a rod monochromat.
Investigative Ophthalmology & Visual Science **21**, 877-879.
 - **Skottun, B. C., Nordby, K. & Rosness, R. (1982).**
Temporal summation in a rod monochromat.
Vision Research **22**, 491-493.

- **Stabell, B., Nordby, K. & Stabell, U. (1987).**
Light-adaptation of the human rod system.
Clinical Vision Sciences **2**, 83-91.
- **Stabell, U., Stabell, B. & Nordby, K. (1986).**
Dark-adaptation of the human rod system: A new hypothesis.
Scandinavian Journal of Psychology **27**, 175-183.
- **Stabell, B., Stabell, U. & Nordby, K. (1986).**
Dark-adaptation in a rod monochromat:
Effect of stimulus size, exposure time and retinal eccentricity.
Clinical Vision Sciences **1**, 75-80.

Zuletzt geändert am 15.2.1996.

© for text by K. Nordby

© for layout by [Sebastian Bonhag](#)

People with online papers in philosophy

Compiled by [David Chalmers](#)

This is a list of individuals who have made available online papers in philosophy and related areas. This practice is very much to be encouraged! Note that this list concentrates mostly on academic philosophers, although some scientists and others in related fields are included. If you know of people I should add to this list, please tell me.

For other sources of online papers see:

- [online papers on consciousness](#)
- [web resources related to consciousness, philosophy, and such](#)
- [directory of online philosophy papers](#) (Joe Lau)

Philosophy of mind (esp. consciousness)

- [Torin Alter](#) (qualia, free will, etc)
- [Tim Bayne](#) (consciousness, personal identity)
- [Ned Block](#) (consciousness, reduction, content, etc)
- [Alex Byrne](#) (consciousness, metaphysics, etc.)
- [Peter Carruthers](#) (consciousness, language)
- [David Chalmers](#) (philosophy of mind and language, cognitive science, metaphysics)
- [Austen Clark](#) (color, consciousness, vision)
- [David Cole](#) (consciousness, language)
- [Craig DeLancey](#) (emotion, consciousness)
- [Daniel Dennett](#) (philosophy of mind)
- [Shaun Gallagher](#) (phenomenology, self-consciousness)
- [Brie Gertler](#) (consciousness, introspection, metaphysics)
- [Benj Hellie](#) (consciousness, Nietzsche)
- [Ted Honderich](#) (consciousness, [free will](#), ethics)
- [Susan Hurley](#) (consciousness, perception, political philosophy)
- [Sean Kelly](#) (perceptual experience, action)
- [Peter Lloyd](#) (consciousness, idealism)
- [Eric Lormand](#) (consciousness, meaning, cognitive architecture)
- [William Lycan](#) (consciousness, epistemology)
- [Thomas Metzinger](#) (consciousness)
- [Gerard O'Brien](#) (consciousness, connectionism)
- [David Papineau](#) (consciousness, epistemology, philosophy of science)

- [Tom Polger](#) (consciousness, evolution)
- [Gregg Rosenberg](#) (*A Place for Consciousness*, etc.)
- [David Rosenthal](#) (consciousness, intentionality, [Sellars](#))
- [William Seager](#) (consciousness, metaphysics)
- [John Searle](#) (consciousness, AI; and [here](#))
- [Pär Sundström](#) (consciousness)
- [Nigel Thomas](#) (imagery, consciousness, [zombie killer](#))
- [Michael Tye](#) (consciousness, qualia, content)

Philosophy of mind (esp. artificial intelligence and cognitive science)

- [William Bechtel](#) (foundations of connectionism & cognitive science)
- [Istvan Berkeley](#) (connectionism, foundations of cognitive science)
- [Mark Bickhard](#) (representation, computation, etc.)
- [Selmer Bringsjord](#) (philosophy of AI, etc.)
- [David Buller](#) (evolution, AI, metaphysics)
- [Ron Chrisley](#) (connectionism, computation, non-conceptual content)
- [Jack Copeland](#) (computation, AI, logic)
- [Hubert Dreyfus](#) (philosophy of AI and technology, Heidegger, etc)
- [Chris Eliasmith](#) (dynamic systems, philosophy of science)
- [Rick Grush](#) (foundations of cognitive science, perceptual content)
- [Larry Hauser](#) (consciousness, AI, [philosophy comics](#), etc.)
- [Brian Keeley](#) (philosophy of neurobiology, artificial life)
- [J. R. Lucas](#) (mechanism, [Gödel](#) and AI, time, etc.)
- [Gregory Mulhauser](#) (robots, consciousness)
- [William Rapaport](#) (foundations of AI)
- [Eric Schwitzgebel](#) (theory of mind, belief, consciousness)
- [Aaron Sloman](#) (foundations of AI, [rock consciousness](#))
- [Stephen Stich](#) (theory of mind, rationality, etc)
- [John Sutton](#) (memory)
- [Tim van Gelder](#) (philosophy of mind, dynamic systems)

Philosophy of mind (miscellaneous)

- [Colin Allen](#) (animal cognition, evolution, *Species of Mind*, etc.)
- [Murat Aydede](#) (language of thought, connectionism, content)
- [John Beloff](#) (mind-body problem, parapsychology)
- [Andrew Brook](#) (Kant, cognitive science, psychoanalysis)
- [Filip Buekens](#) (supervenience, observation)

- [Samir Chopra](#) (reasoning, belief, physics)
- [Adrian Cussins](#) (content, objectivity)
- [Ronnie de Sousa](#) (moral psychology, philosophy of mind & biology)
- [Matthew Elton](#) (consciousness, artificial life, etc.)
- [Peter Gardenfors](#) (concepts, belief revision, [festschrift](#), etc.)
- [Jay Garfield](#) (philosophy of mind, Buddhist thought)
- [Robert Gordon](#) (folk psychology, simulation)
- [York Gunther](#) (nonconceptual content, emotions)
- [David Hilbert](#) (color, philosophy of perception)
- [Jim Hopkins](#) (psychoanalysis, consciousness, interpretation)
- [Steven Horst](#) (metaphysics of mind, consciousness, computationalism)
- [Daniel Hutto](#) (consciousness, idealism)
- [Brendan Lalor](#) (mental content, metaphysics of mind)
- [Stephen Laurence](#) (concepts, content, language)
- [Barry Loewer](#) (philosophy of mind, philosophy of physics)
- [Kirk Ludwig](#) (philosophy of mind, epistemology)
- [Pete Mandik](#) (philosophy of mind, neuroscience, [objectivity](#))
- [Ron McClamrock](#) (philosophy of mind))
- [Ruth Millikan](#) (philosophy of mind and language, philosophy of biology)
- [Barbara Montero](#) (metaphysics of mind, ethics)
- [Thomas Nagel](#) (philosophy of mind, ethics)
- [David Newman](#) (philosophy of mind, chaos theory)
- [Alva Noë](#) (perception, consciousness)
- [Katarzyna Paprzycka](#) (philosophy of action and mind, etc)
- [Teed Rockwell](#) (philosophy of mind)
- [Sarah Sawyer](#) (externalism, self-knowledge)
- [Daniel Stoljar](#) (philosophy of mind, metaphysics)
- [Adam Vinuesa](#) (consciousness, concepts)
- [Wayne Wright](#) (content, perception, Kant)
- [Tad Zawidzki](#) (mental content, philosophy of biology)

Philosophy of language

- [Peter Alward](#) (philosophy of language & mind)
- [Kent Bach](#) (philosophy of language & mind)
- [David Braun](#) (philosophy of language, belief, causation)
- [Harold Brown](#) (concepts, Sellars)
- [Paul Boghossian](#) (externalism, analyticity)
- [Mark Crimmins](#) (belief semantics, philosophy of language & mind)

- [Graeme Forbes](#) (belief semantics, metaphysics)
- [Christopher Gauker](#) (philosophy of language & mind)
- [Delia Graff](#) (philosophy of language)
- [Mitchell Green](#) (philosophy of language, etc.)
- [Richard Heck](#) (philosophy of language & logic, mental content)
- [Richard Holton](#) (philosophy of language, metaphysics, ethics)
- [Richard Horsey](#) (philosophy of language, linguistics)
- [John Humphrey](#) ([Kripke's Wittgenstein](#))
- [Henry Jackman](#) (philosophy of language and mind, William James)
- [Frank Jackson](#) (philosophy of language, ethics)
- [Bjorn Jespersen](#) (philosophy of language, logic)
- [Gary Kemp](#) (truth, consciousness)
- [Jeffrey King](#) (philosophy of language)
- [Joe Lau](#) (belief semantics, [mind/cogsci resources](#))
- [Ernest Lepore](#) (philosophy of language and mind)
- [Peter Ludlow](#) (belief semantics, mental content, cyberspace)
- [Michael Nelson](#) (philosophy of language, time)
- [Pavel Materna](#) (semantics, intensional logic)
- [Michael Pelczar](#) (philosophy of language, metaphysics)
- [Carlo Penco](#) (philosophy of language, AI)
- [Jaroslav Peregrin](#) (philosophy of language and logic)
- [John Perry](#) (philosophy of language & mind)
- [Jennifer Saul](#) (philosophy of language, feminism)
- [Dan Sperber](#) (language, cognition, culture)
- [Jason Stanley](#) (philosophy of language, metaphysics)
- [Richmond Thomason](#) (philosophy of language, reasoning)
- [Howard Wettstein](#) (philosophy of language, philosophy of religion)
- [Takashi Yagisawa](#) (philosophy of language)

Metaphysics

- [Gordon Barnes](#) (modality, philosophy of mind)
- [Karen Bennett](#) (metaphysics)
- [John Bigelow](#) (metaphysics, epistemology, etc.)
- [Alexander Bird](#) (metaphysics, epistemology, philosophy of science)
- [Simon Blackburn](#) (causation, reviews)
- [Jonathan Cohen](#) (color, philosophy of language)
- [Andrew Chrucky](#) (metaphysics, [Wilfrid Sellars archive](#))

- [Cian Dorr](#) (metaphysics)
- [Crawford Elder](#) (metaphysics, philosophy of mind)
- [Michael Fara](#) (dispositions, epistemic logic)
- [Kit Fine](#) (metaphysics, logic)
- [Thomas Hofweber](#) (metaphysics, epistemology, philosophy of language and logic)
- [Tomis Kapitan](#) (free will, philosophy of language and mind, indexicality)
- [Peter King](#) (metaphysics, etc.)
- [Ned Markosian](#) (metaphysics)
- [Neil McKinnon](#) (metaphysics)
- [D.H. Mellor](#) (philosophy of time, etc)
- [Peter Menzies](#) (causation, etc)
- [Mark Moyer](#) (identity, supervenience)
- [Laurie Paul](#) (metaphysics)
- [Robert Pendleton](#) (time, free will)
- [John Post](#) (metaphysics, epistemology)
- [Allan Randall](#) (metaphysics, QM, history of philosophy)
- [Theodore Sider](#) (metaphysics, etc.)
- [Brock Sides](#) (metaphysics, history)
- [Barry Smith](#) (ontology)
- [Quentin Smith](#) (time, cosmology, language, religion, etc)
- [Norman Swartz](#) (metaphysics, epistemology)
- [David Thompson](#) (metaphysics, phenomenology, intentionality)
- [Kent van Cleave](#) (evolutionary metaphysics, etc)
- [Achille Varzi](#) (events, vagueness, topology, metaphysics)
- [Brian Weatherson](#) (M&E, language, metaphilosophy, decision theory)
- [V. Alan White](#) (free will, twin paradox, [philosophy songs](#))
- [Jessica Wilson](#) (supervenience, apriority)
- [Stephen Yablo](#) (metaphysics, philosophy of mind)
- [Ed Zalta](#) (metaphysics, philosophy of language)

Epistemology

- [Tim Black](#) (skepticism, contextualism)
- [William Boardman](#) (epistemology, metaphysics, logic, law)
- [Laurence Bonjour](#) (epistemology, philosophy of mind)
- [Albert Casullo](#) (a priori knowledge, epistemology)
- [Keith DeRose](#) (epistemology, philosophy of religion)
- [Richard Feldman](#) (epistemology)
- [Gilbert Harman](#) (epistemology, ethics, mind and language)

- [David Henderson](#) (epistemology, philosophy of science)
- [Christopher Hookway](#) (epistemology, pragmatism)
- [Michael Huemer](#) (epistemology, metaphysics, etc.)
- [Nenad Miscevic](#) (intuitions, semantics, political philosophy)
- [Adam Morton](#) (epistemology, ethics, language)
- [Ram Neta](#) (skepticism, contextualism, knowledge)
- [John Pollock](#) (reasoning, AI, decision theory)
- [James Pryor](#) (epistemology, metaphysics)
- [Wolfgang Spohn](#) (reasoning, probability, causation, philosophy of science, etc)
- [Matthias Steup](#) (epistemology)
- [Crispin Wright](#) (epistemology, metaphysics, philosophy of mathematics)

Philosophy of science

- [Andre Ariew](#) (philosophy of biology)
- [Yuri Balashov](#) (philosophy of physics, metaphysics)
- [Mark Bedau](#) (philosophy of biology, mind, economics, complex systems)
- [Gordon Belot](#) (philosophy of physics)
- [Nick Bostrom](#) (doomsday/anthropic reasoning, probability, transhumanism, etc)
- [Richard Burian](#) (philosophy of biology, philosophy of science)
- [Jeremy Butterfield](#) (philosophy of physics)
- [Craig Callender](#) (philosophy of physics)
- [Rob Clifton](#) (philosophy of physics)
- [John Collier](#) (philosophy of science)
- [Stephen Downes](#) (philosophy of biology)
- [Malcolm Forster](#) (causation, [simplicity](#))
- [Ken Gemes](#) (philosophy of science, Nietzsche)
- [Peter Godfrey-Smith](#) (philosophy of biology)
- [Adolf Grunbaum](#) (philosophy of cosmology, psychoanalysis)
- [William Harms](#) (philosophy of biology)
- [Gary Hardcastle](#) (philosophy of science, [Monty Python](#))
- [Richard Healey](#) (philosophy of physics)
- [Philip Kitcher](#) (philosophy of science, epistemology, ethics)
- [Alexander Levine](#) (philosophy of science, metaphysics)
- [David Malament](#) (philosophy of physics)
- [Huw Price](#) (philosophy of physics, metaphysics, [time's arrow](#))
- [Greg Ransom](#) (philosophy of science)
- [Simon Saunders](#) (philosophy of physics)

- [Oliver Schulte](#) (philosophy of science, epistemology)
- [Elliott Sober](#) (philosophy of biology, philosophy of science)
- [Laszlo Szabo](#) (foundations of quantum mechanics)
- [Paul Teller](#) (philosophy of science, metaphysics)
- [Paul Thagard](#) (philosophy of science, cognitive science)
- [Philip Thonemann](#) (philosophy of science)
- [Bas van Fraassen](#) (philosophy of science, metaphysics, etc.)

Philosophy of logic and mathematics

- [Varol Akman](#) (philosophy of logic & language, situation theory)
- [Jeremy Avigad](#) (logic, philosophy of mathematics)
- [Timothy Bays](#) (philosophy of logic)
- [J.C. Beall](#) (philosophy of logic)
- [Nuel Belnap](#) (philosophy of logic, agency, time)
- [Andrew Boucher](#) (philosophy of mathematics, computation)
- [Mark Colyvan](#) (philosophy of mathematics)
- [John Etchemendy](#) (philosophy of logic, computation)
- [Solomon Feferman](#) (logic, philosophy of mathematics)
- [Melvin Fitting](#) (modal logic, philosophical logic)
- [Luciani Floridi](#) (philosophy of computing, AI, ethics)
- [Torkel Franzen](#) (philosophy of mathematics)
- [Michael Glanzberg](#) (philosophy of logic and language)
- [William Greenberg](#) (identity, logic, reference)
- [Patrick Grim](#) (philosophy of logic and mathematics)
- [Richard Jeffrey](#) (probabilistic thinking)
- [Gert-Jan Lokhorst](#) (logic, Wittgenstein, philosophy of mind)
- [Penelope Maddy](#) (philosophy of mathematics)
- [Gary Mar](#) (philosophy of mathematics, philosophy of religion)
- [John McFarlane](#) (philosophy of logic, M&E)
- [Mika Oksanen](#) (philosophical logic, metaphysics)
- [Lorenzo Pena](#) (philosophy of logic, metaphysics and epistemology, ethics)
- [Agustin Rayo](#) (philosophy of logic)
- [Greg Restall](#) (philosophical logic, metaphysics)
- [Nick Smith](#) (paradoxes, logic, time travel)
- [Roy Sorenson](#) (paradoxes, philosophy of logic and language, metaphysics and epistemology)
- [Peter Suber](#) (all sorts, e.g. paradoxes, [Nomic](#), software)
- [Jamie Tappenden](#) (philosophy of logic and mathematics, Frege)
- [Neil Tennant](#) (philosophy of logic and language, philosophy of science)

- [Richard Zach](#) (philosophy of mathematics)

Philosophy of religion

- [Marilyn Adams](#) (philosophy of religion)
- [William Alston](#) (philosophy of religion, epistemology)
- [David Basinger](#) (philosophy of religion)
- [William Lane Craig](#) (philosophy of religion, cosmology)
- [Richard Davis](#) (philosophy of religion, ethics)
- [Steven Davis](#) (philosophy of religion)
- [Alfred Freddoso](#) (philosophy of religion, history of philosophy)
- [Daniel Howard-Snyder](#) (philosophy of religion, epistemology)
- [Robert Koons](#) (philosophy of religion, metaphysics)
- [Graham Oppy](#) (philosophy of religion, cosmology)
- [Alvin Plantinga](#) (philosophy of religion, metaphysics)
- [Nicholas Rescher](#) (philosophy of religion)
- [J. Wesley Robbins](#) (religious naturalism, pragmatism)
- [James F. Sennett](#) (philosophy of religion, freedom, abortion, etc)
- [Quentin Smith](#) (philosophy of cosmology, philosophy of religion)
- [Michael Sudduth](#) (philosophy of religion)
- [Richard Swinburne](#) (philosophy of religion)
- [Peter van Inwagen](#) (philosophy of religion)
- [Edward Wierenga](#) (philosophy of religion)

Value theory (ethics, social/political philosophy, aesthetics)

- [Carla Bagnoli](#) (ethics)
- [Gregory Currie](#) (aesthetics, philosophy of mind)
- [Ronald Dworkin](#) (moral philosophy)
- [William Edmundson](#) (philosophy of law)
- [Jon Elster](#) (social/political philosophy, rationality)
- [David Estlund](#) (moral & political philosophy)
- [Patricia Greenspan](#) (ethics, emotion, free will)
- [John Hardwig](#) (applied ethics, epistemology)
- [Sally Haslanger](#) (feminism, race, metaphysics)
- [Dan Haybron](#) (happiness, evil)
- [Bennett Helm](#) (moral psychology)
- [Dale Jamieson](#) (environmental ethics)
- [Mark Kalderon](#) (meta-ethics, philosophy of mind)

- [Hugh LaFollette](#) (ethics)
- [Michelle Mason](#) (ethics, moral psychology)
- [Diana Tietjen Meyers](#) (feminism, ethics)
- [Elijah Millgram](#) (practical reasoning, Hume)
- [Louis Pojman](#) (equality, philosophy of religion)
- [Joseph Raz](#) (philosophy of law)
- [William Ruddick](#) (medical ethics, etc)
- [David Schmidtz](#) (ethics, social philosophy)
- [David Sobel](#) (well-being, rationality)
- [Alan Soble](#) (philosophy of sex, love, & gender)
- [Robert Streiffer](#) (ethics, M&E)
- [Peter Unger](#) (ethics)
- [Julie Van Camp](#) (aesthetics)
- [Rob van Gerwen](#) (aesthetics, Kant)
- [J. David Velleman](#) (practical reason)
- [R. Jay Wallace](#) (ethics, practical reasoning)
- [Ralph Wedgwood](#) (meta-ethics)
- [Nick Zangwill](#) (ethics, aesthetics, metaphysics)

History of philosophy

- [Julia Annas](#) (ancient Greek philosophy, ethics)
- [Andrew Carpenter](#) (Kant, philosophy of language and mind)
- [Alan Fox](#) (Chinese philosophy)
- [Cynthia Freeland](#) (ancient philosophy, aesthetics, feminism)
- [Chad Hansen](#) (Chinese philosophy)
- [Lars Hertzberg](#) (Wittgenstein, moral psychology)
- [Paul Hoffman](#) (Leibniz, free will)
- [Andrew Irvine](#) (Russell, analytic philosophy)
- [John Kilcullen](#) (medieval philosophy)
- [Gyula Klima](#) (medieval philosophy)
- [Justin Leiber](#) (Wittgenstein, philosophy of mind, etc.)
- [Kurt Mosser](#) (Kant, epistemology)
- [Andrew Lilico](#) (Wittgenstein, metaphysics)
- [Lex Newman](#) (Descartes, Locke)
- [Stephen Palmquist](#) (Kant, etc.)
- [Joseph Ransdell](#) (Peirce)
- [Jay Rosenberg](#) (history, metaphysics, epistemology)

- [Kelley Ross](#) (all sorts)
- [Paul Vincent Spade](#) (mediaeval philosophy)
- [Saul Traiger](#) (Hume, philosophy of mind)
- [Allen Wood](#) (Kant, ethics, history of philosophy)

"Philosophical" cognitive scientists

- [Bernard Baars](#) (consciousness, *[A Cognitive Theory of Consciousness](#)*)
- [Daryl Bem](#) (parapsychology, sexual orientation)
- [Susan Blackmore](#) (memes, parapsychology)
- [Rodney Brooks](#) (robotics)
- [William Calvin](#) (neuroscience, evolution)
- [Axel Cleeremans](#) (implicit learning, connectionism, consciousness)
- [Rodney Cotterill](#) (consciousness, neuroscience)
- [Stan Franklin](#) (computational models of consciousness)
- [Robert French](#) (foundations of AI, representation)
- [Christopher Green](#) (foundations of cognitive science)
- [Stevan Harnad](#) (foundations of cognitive science)
- [David Israel](#) (foundations of AI, semantics)
- [Steven Lehar](#) (perception, gestalt theory, consciousness)
- [Ron Loui](#) (reasoning, foundations of AI)
- [Bruce MacLennan](#) (connectionism, [consciousness](#), etc.)
- [John McCarthy](#) (foundations of AI)
- [Marvin Minsky](#) (foundations of AI)
- [Hans Moravec](#) (AI & robotics)
- [Kevin O'Regan](#) (perception, qualia, change blindness)
- [Daniel Osherson](#) (reasoning, vagueness)
- [Jordan Pollack](#) (connectionism, dynamic systems)
- [Eyal Reingold](#) (unconscious perception, implicit memory)
- [Gerd Sommerhoff](#) (consciousness)
- [Ron Sun](#) (connectionism, symbols, consciousness)
- [Charles Tart](#) (consciousness, transpersonal psychology)
- [John Taylor](#) (consciousness, neural networks)
- [Francisco Varela](#) (phenomenology, neuroscience)
- [Max Velmans](#) (consciousness)

"Philosophical" physicists & mathematicians

- [Gregory Chaitin](#) (foundations of math, algorithmic information theory)

- [John Cramer](#) (physics, [science columns](#), [transactional interpretation](#) of QM)
- [David Deutsch](#) (foundations of quantum mechanics)
- [Harvey Friedman](#) (foundations of mathematics, set theory)
- [Sheldon Goldstein](#) (foundations of QM, esp. Bohm's interpretation)
- [John Gribbin](#) (physics, etc.)
- [Stuart Hameroff](#) (QM & consciousness)
- [Andrew Hodges](#) (twistor theory, [Alan Turing](#))
- [Piet Hut](#) (consciousness, astrophysics, computation)
- [Jacques Mallah](#) (foundations of QM)
- [Don Page](#) (foundations of QM, consciousness)
- [Jack Sarfatti](#) ("physics/consciousness research group" & more)
- [Alan Sokal](#) (physics, [parody](#); see also [the Sokal affair](#))
- [Henry Stapp](#) (foundations of QM, consciousness)
- [Victor Stenger](#) (foundations of QM, skepticism)

Other interesting individuals

- [Bill Adams](#) (consciousness)
- [Tom Clark](#) (naturalism, free will, consciousness, etc)
- [Richard Dawkins](#) (evolutionary biology)
- [Christian DeQuincey](#) (consciousness, spirituality)
- [Philip Dorrell](#) (consciousness, evolution, philosophy)
- [Eric Drexler](#) (nanotechnology)
- [Bruce Edmonds](#) (complexity, holism)
- [Claus Emmeche](#) (artificial life, philosophy of biology)
- [Leon Felkins](#) (paradoxes, [social dilemmas](#), political philosophy)
- [Liane Gabora](#) (evolution of ideas, creativity, consciousness)
- [James Gleick](#) (technology, science)
- [George Johnson](#) (science journalism)
- [Philip Johnson](#) (evolution, creationism)
- [Roger B. Jones](#) (lots of philosophy, science, mathematics, etc.)
- [Jaron Lanier](#) (virtual reality, philosophy of AI, [more zombies](#))
- [Greg Nixon](#) (consciousness, education)
- [Howard Pattee](#) (systems science)
- [David Pearce](#) ("HedWeb"; lots of good stuff, e.g. [Conscious Mind review](#))
- [Cosma Shalizi](#) (science, philosophy, everything else)
- [Ken Wilber](#) (integral psychology)
- [Robert M. Young](#) (philosophy of science, psychoanalysis)
- [Eliezer Yudkowsky](#) (AI, technology, singularity)

Go to:

- [David Chalmers' home page](#)

Web resources related to consciousness, philosophy, and such.

Compiled by [David Chalmers](#)

Here are a small number of high-quality academic resources on the web that I find useful or interesting. The emphasis is on sites containing real intellectual content.

See also my lists of [people with online papers in philosophy](#) and of [online papers on consciousness](#).

Consciousness resources

- [Association for the Scientific Study of Consciousness](#)
- [Brain Project](#)
- [Center for Consciousness Studies \(Arizona\)](#)
- [Consciousness in the Natural World Project](#)
- [Journal of Consciousness Studies](#)
- [Metaphysics of Consciousness](#)
- [New Scientist on consciousness](#)
- [NYU research seminar on consciousness](#)
- [Online papers on consciousness](#)
- [PSYCHE: An interdisciplinary journal for research on consciousness](#)
- [PSYCHE-D mailing list archives](#)
- [Zombies on the Web](#)

Philosophy resources

- [Aesthetics Online](#)
- [American Philosophical Association](#)
- [ANU Working Papers in Social and Political Philosophy](#)
- [BEARS: Reviews in Moral and Political Philosophy](#)
- [CogPrints Philosophy Papers](#)
- [Dictionary of Philosophical Terms](#)
- [Dictionary of Philosophy of Mind](#)
- [Directory of Online Philosophy Papers](#)
- [Epistemology Research Guide](#)
- [Field Guide to the Philosophy of Mind](#)
- [FOLDOP: Free Online Dictionary of Philosophy](#)

- [Folk Psychology vs. Mental Simulation](#)
- [Formal Ontology](#)
- [International Philosophical Preprint Exchange](#)
- [Internet Encyclopedia of Philosophy](#)
- [Mediaeval Logic and Philosophy](#)
- [Metaphysics Research Lab](#)
- [NYU Philosophy Papers](#)
- [Paideia Archive](#)
- [Philosophers' Web Magazine](#)
- [Philosophy Books Online](#)
- [Philosophy of Science Archive](#)
- [Stanford Encyclopedia of Philosophy](#)
- [University of Chicago Philosophy Project](#)
- [Washington University PNP Archive](#)

Philosophy journals online

- [ANALYSIS papers](#)
- [Electronic Journal of Analytic Philosophy](#)
- [HYLE: Journal of the Philosophy of Chemistry](#)
- [JSTOR journals](#)
 - [Ethics](#)
 - [Journal of Philosophy](#)
 - [Mind](#)
 - [Nous](#)
 - [Philosophical Perspectives](#)
 - [Philosophical Quarterly](#)
 - [Philosophical Review](#)
 - [Philosophy and Phenomenological Research](#)
 - [Philosophy and Public Affairs](#)
- [Metaphysical Review](#)
- [Nordic Journal of Philosophical Logic](#)
- [Sorites: Electronic Quarterly of Analytic Philosophy](#)
- [Philosophy, Psychiatry, and Psychology](#)
- [STOA: Undergraduate Journal of Philosophy](#)
- [Truth](#)

Cognitive science resources

- [Artificial Life: Online Papers](#)
- [Behavioral and Brain Sciences Index](#)
- [Classics in the History of Psychology](#) (and [more](#))
- [Cognitive Science Article Archive](#)
- [Cognitive Science of Metaphor Online](#)
- [CogPrints: Cognitive Science E-Print Archive](#)
- [Human-Nature.com](#) (inc. Darwin, evolutionary psychology)
- [Mind and Body: Rene Descartes to William James](#)
- [MIT Encyclopedia of Cognitive Science](#)
- [Neuroprose archive](#) (ftp)
- [Psychological Online Documents](#)
- [PSYCOLOQUY](#)
- [The Joy of Visual Perception](#)

Meta-resources

- [Guide to Philosophy on the Internet](#) (Peter Suber)
- [HIPPIAS: Philosophy search engine](#)
- [Meta-Encyclopedia of Philosophy](#) (Andrew Chrucky)
- [Mind/brain resources](#) (Valerie Hardcastle)

Disorders of consciousness

- [Blindsight Learning Center](#)
- [Face Blind!](#) (and [here](#))
- [Mindblindness and Asberger's Syndrome](#)
- [Multiple Personality Disorder](#) (and [here](#) and [here](#))
- [Split Brain Consciousness](#)
- [Synesthesia](#) (and [here](#))
- [Visual Experience in a Complete Achromat](#)
- [Seeing with Sound](#)

Interactive demonstrations

- [Binocular Rivalry](#) (and other [visual effects](#))
- [Brain and Behavior](#)
- [Change Blindness](#) (and [here](#))
- [Fun Things in Visual Perception](#)
- [IllusionWorks Home Page](#)

- [Interactive Illustrations of Color Perception](#)
- [Live Artificial Life Page](#)
- [Newcomb's Problem](#)
- [Philosophical Toys](#)
- [Physics 2000](#)
- [Sodaplay](#)
- [Whole Brain Atlas](#)

Miscellaneous sites with interesting content

- [Alan Turing Home Page](#)
- [Anthropic Principle Archive](#)
- [Artificial Self-Replication Page](#)
- [Economics Working Papers](#)
- [Foresight Exchange](#)
- [KurzweilAI.net](#)
- [Memetics](#)
- [Mind Uploading Home Page](#)
- [Physics E-Print Archive](#)
- [Singularity Institute](#)
- [The Edge \(Science and Knowledge\)](#)
- [The Headless Way](#)
- [Thought Experiments](#)
- [Turing Test home page](#)

Go to:

- [David Chalmers' home page](#)

Philosophy 596V: Philosophy and Cognitive Science

Foundational Issues in the Science of Consciousness

David Chalmers

E-mail: chalmers@arizona.edu

Phone: 621-7105

Web: <http://www.u.arizona.edu/~chalmers/>

Office hours: Thursday 1:30-3pm (Social Sciences 226A)

Class meeting (Fall 1999): Tuesday 3:30-6pm, Social Sciences 311.

Overview

In recent years there has been an explosion of work on consciousness, within philosophy, psychology, and neuroscience. This seminar will examine the science of consciousness, addressing both empirical and philosophical issues within it.

Some empirical areas to be covered include:

- the search for the neural correlate of consciousness
- neuropsychological disorders of consciousness
- the processes involved in conscious vision
- implicit processes in memory, learning, and perception
- the relationship between consciousness and attention
- cognitive models of consciousness
- consciousness and metacognition
- animal consciousness
- phenomenological approaches to consciousness.

Some philosophical issues to be addressed include:

- What are the varieties of "consciousness"?
- Can consciousness be explained in terms of neurobiology?
- How can we gain experimental access to consciousness?
- What is an appropriate epistemology for the science of consciousness?
- What might a final theory of consciousness look like?
- What can one being know about the consciousness of another?
- Might consciousness differ radically between individuals?

Some weeks will be devoted to "empirical" topics and some weeks to "philosophical" topics, but for any given topic we will read papers by both empirical researchers and philosophers, often spending half the meeting on each. In the empirical areas, we will make an effort to address the foundational questions that lie behind the research. In the philosophical areas, we will examine how empirical work can illuminate and occasionally reshape these issues.

This course is intended as an interdisciplinary seminar. Material in both philosophy and cognitive science will be approached in such a way to be accessible to students without much background in one or the other. Any graduate student in philosophy, psychology, or other relevant areas would be welcome.

Readings

There is no required textbook. Recommended texts are:

Block, Flanagan, and Guzeldere (eds.): *The Nature of Consciousness: Philosophical Debates*. MIT Press, 1997.

Cohen and Schooler (eds.): *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

Many readings will be available on the web, and others will be made available for copying. Students will be expected to make brief seminar presentations of the readings.

Web page

The web page for the class is at

<http://www.u.arizona.edu/~chalmers/phil596v.html>

Mailing list

I will set up a mailing list, scicon@listserv.arizona.edu, for class discussion. Everyone will be expected to make reasonably regular contributions to this list (at least one reasonably substantial posting every two weeks), discussing issues arising from the readings, from class discussion, and from the mailing list itself. Think of this as a substitute for biweekly short papers of 1-2 pages each.

Assessment

Assessment will be based most heavily on a final paper, and will also be based on in-class presentations, mailing list contributions, and class participation.

Schedule

Here is a very approximate week-by-week plan for the course, with associated readings. Note that this is very likely to be revised and supplemented as things develop. Readings that are not available on the web will be made available in the department office. Full bibliographical information can be found in my online [bibliographies](#) of consciousness in philosophy and of consciousness in science.

1. What is consciousness?

- Norton Nelkin, "What is consciousness?"
- Edoardo Bisiach, "The (haunted) brain and consciousness".

[Week 1 discussion](#)

2. Can consciousness be reductively explained?

- David Chalmers, [Facing up to the problem of consciousness](#)
- Paul Churchland, "The rediscovery of light".
- Daniel Dennett, [Facing backwards on the problem of consciousness](#)
- Francis Crick and Christof Koch, [Why neuroscience may be able to explain consciousness](#)
- David Chalmers, [Moving forward on the problem of consciousness](#)

[Week 2 discussion](#)

3. What can be known about the consciousness of another?

- Frank Jackson, [Epiphenomenal qualia](#)
- Paul Churchland, "Reduction, qualia, and the direct introspection of brain states"
- Evan Thompson, "Novel colours"
- Knut Nordby, [Vision in a complete achromat: A personal account](#)

[Week 3 discussion](#)

4. Neural correlates of consciousness I.

- Francis Crick and Christof Koch, "Toward a neurobiological account of consciousness"
- Joseph Bogen, "On the neurophysiology of consciousness, part I: An overview."
- Ned Block, [How to find the neural correlate of consciousness](#)
- David Chalmers, [On the search for the neural correlate of consciousness](#)

[Week 4 discussion](#)

5. Neural correlates of consciousness II.

- D.L. Sheinberg & Nikos Logothetis, [The role of temporal cortical areas in perceptual organization](#)
- Francis Crick and Christof Koch, [Consciousness and neuroscience](#)
- Davida Teller and E.N. Pugh Jr., Linking propositions in color vision
- David Chalmers, [What is a neural correlate of consciousness?](#)

[Week 5 discussion](#)

6. Blindsight

- Petra Stoerig and Alan Cowey, Blindsight in man and monkey
- Semir Zeki and Dominic ffytche, The Riddoch syndrome: Insights into the neurobiology of conscious vision
- Ned Block, [On a confusion about a function of consciousness](#)

[Week 6 discussion](#)

7. Conscious and unconscious perception

- David Milner and Melvyn Goodale, [The visual brain in action \(precis\)](#)
- Philip Merikle & Meredyth Daneman, [Psychological investigations of unconscious perception](#)
- Anthony Greenwald, New Look 3: Unconscious cognition reclaimed
- Max Velmans, [Is human information processing conscious?](#)

[Week 7 discussion](#)

8. Change blindness, consciousness, and attention.

- Arien Mack and Irvin Rock, [Inattention blindness: An overview](#)
- Daniel Simons, Change blindness
- Jeremy Wolfe, [Inattention amnesia](#)
- Valerie Hardcastle, [Consciousness and attention: A distinction with a difference](#)
- Alva Noe and Evan Thompson, [Beyond the grand illusion: What change blindness really teaches us about vision](#)

[Week 8 discussion](#)

9. Cognitive models of consciousness.

- Bernard Baars, [In the theatre of consciousness](#)
- Daniel Schacter, On the relation between memory and consciousness: Dissociable interactions and conscious experience
- Daniel Dennett & Marcel Kinsbourne, [Time and the observer: The where and when of consciousness in the brain](#)

[Week 9 discussion](#)

10. Consciousness and metacognition.

- Thomas Nelson, Consciousness and metacognition
- Marcia Johnson and John Reeder, Consciousness as meta-processing
- David Rosenthal, [Consciousness, content, and metacognitive judgments](#)

[Week 10 discussion](#)

11. Animal consciousness.

- Gordon Gallup, [Are animals self-aware? Yes](#)
- Daniel Povinelli, [Are animals self-aware? Maybe not](#)
- Thomas Nagel, [What is it like to be a bat?](#)
- Kathleen Akins, What is it like to be boring and myopic?

[Week 11 discussion](#)

12. The inverted spectrum.

- Stephen Palmer, [Color, consciousness, and the isomorphism constraint](#)
- C.L. Hardin, Color quality and color structure
- Martine Nida-Rumelin, Pseudonormal vision: An actual case of spectrum inversion?

13. Epistemology of consciousness.

- Eyal Reingold and Philip Merikle, Theory and measurement in the study of unconscious processes
- Jonathan Schooler, Consciousness and the limits of language
- Alvin Goldman, Can science know when you're conscious? Epistemological foundations of consciousness research

14. First-person approaches.

- David Chalmers, [First-person methods in the science of consciousness](#)
- Francisco Varela, [Neurophenomenology: A methodological remedy for the hard problem](#)
- Steven Horst, [Phenomenology and psychophysics](#)

15. To be determined

Web resources

Some useful resources on the web include:

- [Online papers on consciousness](#)
- [Consciousness in Philosophy: A Bibliography](#)
- [Consciousness in Science: A Bibliography](#)
- [Web resources related to consciousness, philosophy, and such](#)

Philosophy 596B: Metaphysics

Meaning, Reason, and Possibility

[David Chalmers](#)

E-mail: chalmers@arizona.edu

Web: <http://www.u.arizona.edu/~chalmers/>

Office hours (Spring 2002): Wednesday 1:30-3pm (Social Sciences 226A)

Class meetings: Monday/Wednesday 3:30-5pm, Social Sciences 311.

Overview

This seminar is nominally listed as metaphysics, but it will address issues in the philosophy of language, metaphysics, epistemology, and the philosophy of mind. The seminar will explore the constitutive connections among reason, meaning, and possibility.

Frege's notion of sense tied meaning to reason. Carnap's notion of intension tied meaning to possibility. In conjunction with a Kantian view connecting reason and possibility, this yielded a golden triangle of connections among the three notions. Kripke's work broke the triangle by severing meaning and possibility from reason. One might see the central focus of this seminar as the project of once again articulating an approach to meaning and possibility on which they are constitutively tied to reason, thus restoring the golden triangle.

We will be discussing a fair amount of my own recent work on these topics, as well as work by others. Topics to be discussed include:

- Fregean theories of meaning
- two-dimensional semantics
- epistemic possibility and epistemic space
- conceptual analysis and its role in metaphysics
- narrow content and the semantics of thought
- conceivability and possibility
- the mind-body problem

Readings

There is no textbook. Many readings will be available on the web, and other will be made available for photocopying in a folder in the department office.

Web page

The web page for the class is at

<http://www.u.arizona.edu/~chalmers/phil596b.html>

Mailing list

I will set up a mailing list, phil596b@philosophy.arizona.edu, for class discussion. Everyone enrolled in the class is expected to make reasonably regular contributions to this list: at least one reasonably substantial posting every week (you can miss three weeks without penalty), discussing issues arising from the readings, from class discussion, and from the mailing list itself. Of course people are welcome and encouraged to post more often. These postings substitute for biweekly short papers of 1-2 pages each.

Assessment

Assessment will be based most heavily on a final paper, and will also be based on in-class presentations, mailing list contributions, and class participation.

Schedule

Here is a very approximate plan for the course, with associated readings. This is very likely to be revised as things develop.

Note that I will be out of town on quite a few meeting dates: Wed 1/23 (Cornell), Wed 2/6 (Rutgers), Wed 2/20 (ANU), Mon 2/25 (ANU), Wed 3/21 (North Carolina), Mon-Wed 4/8-10 (consciousness conference). These dates will be made up either by having double meetings on neighboring dates, or by rescheduling. I hope to have a more precise schedule of meetings soon.

Meeting 1 (Wed 1/9): Introduction

PART I: SENSE AND INTENSION

Meeting 2 (Mon 1/14): Frege on sense

- Frege, On sense and reference
- Chalmers, [On sense and intension](#), sections 1-3

Meeting 3 (Mon 1/14): Carnap on intension

- Carnap, *Meaning and Necessity*, pp. 1-32
- Chalmers, [The foundations of two-dimensional semantics](#), section 1.1

Meeting 4 (Wed 1/16): Epistemic intensions

- Chalmers, [On sense and intension](#), section 4
- Chalmers, [The foundations of two-dimensional semantics](#), sections 3.1-3.2

Meeting 5 (Tue 1/22, 6:30pm): Indexicality and centering

- Perry, Frege on demonstratives
- Chalmers, [On sense and intension](#), section 5

Meeting 6 (Mon 1/28): Kripke's modal argument

- Kripke, *Naming and Necessity*, Lecture 1 (esp. pp. 39-63)

Meeting 7 (Wed 1/30): Two-dimensional intensions

- Chalmers, [On sense and intension](#), section 6

Meeting 8 (Mon 2/4): Kripke's epistemic argument

- Kripke, *Naming and Necessity*, Lecture 2 (esp. pp. 79-97)
- Chalmers, [On sense and intension](#), section 7

Meeting 9 (Mon 2/4): Variability

- Burge, Sinning against Frege
- Chalmers, [On sense and intension](#), section 8

PART II: CONCEPTUAL ANALYSIS AND THE SCRUTABILITY OF TRUTH

Meeting 10 (Mon 2/11): Conceptual analysis and scrutability

- Chalmers and Jackson, [Conceptual analysis and reductive explanation](#), sections 1-4
- Chalmers, [Does conceivability entail possibility](#), section on "Inscrutabilities"

Meeting 11 (Wed 2/13): Block and Stalnaker against scrutability

- Block and Stalnaker, [Conceptual analysis, dualism, and the explanatory gap](#), sections 1-9
- Chalmers and Jackson, [Conceptual analysis and reductive explanation](#), sections 5-6

Meeting 12 (Mon 2/18): Yablo against scrutability

- Stephen Yablo, [Coulda, woulda, shoulda](#), especially section 20
- Chalmers, [Does conceivability entail possibility](#), section on "modal rationalism and logical empiricism"

Meeting 13 (Mon 2/18): Hard cases

- Chalmers, [Does conceivability entail possibility](#), second half of section on "inscrutabilities"

PART III: APRIORITY AND EPISTEMIC SPACE

Meeting 14 (Wed 2/27): Apriority

- Bonjour, *In Defense of Pure Reason*, Chapter 1

Meeting 15 (Wed 2/27): Quine against apriority Quine, [Two dogmas of empiricism](#)

- Quine, [Two dogmas of empiricism](#)

Meeting 16 (Mon 3/4): Epistemic possibility and epistemic space

- Chalmers, [The nature of epistemic space](#), sections 1-3

Meeting 17 (Wed 3/6): Scenarios as possible worlds

- Chalmers, [The nature of epistemic space](#), section 4
- Lewis, Possible worlds
- Lewis, *On the Plurality of Worlds*, Chapter 3

Meeting 18 (Mon 3/18): Constructing possible worlds

- Chalmers, [The nature of epistemic space](#), section 5

PART IV: TWO-DIMENSIONAL SEMANTICS

Meeting 20 (Mon 3/25): Stalnaker's diagonal construction

- Stalnaker, Assertion
- Block and Stalnaker, [Conceptual analysis, dualism, and the explanatory gap](#), section 10

Meeting 21 (Wed 3/27): The contextual and epistemic understandings

- Chalmers, [The foundations of two-dimensional semantics](#), section 2-4

3. Other approaches

- Kaplan, On the logic of demonstratives
- Evans, Reference and contingency
- Davies and Humberstone, Two notions of necessity
- Chalmers, [The foundations of two-dimensional semantics](#), section 5

PART V: MENTAL CONTENT

1. Arguments for externalism

- Putnam, The meaning of 'meaning'
- Burge, Individualism and the mental

2. Narrow content

- Loar, Social content and psychological content
- Stalnaker, Narrow content
- Horgan and Tienson, The intentionality of phenomenology and the phenomenology of intentionality

3. Epistemic content as narrow content

- Chalmers, [The components of content](#), sections 1-6

4. Belief ascriptions

- Schiffer, The mode-of-presentation problem
- Chalmers, [The components of content](#), section 7

PART VI: CONCEIVABILITY, POSSIBILITY, AND THE MIND-BODY PROBLEM

1. Varieties of conceivability

- Chalmers, [Does conceivability entail possibility](#) (through "GAPS")

2. Does conceivability entail possibility?

- Chalmers, [Does conceivability entail possibility](#) ("GAPS" and "THESES")
- van Cleve, *Conceivability and the Cartesian argument for dualism*
- Yablo, *Is conceivability a guide to possibility?*

3. The mind-body problem

- Chalmers, [Consciousness and its place in nature](#), sections 1-5

4. Two-dimensionalism and the mind-body problem

- Kripke, *Naming and Necessity*, pp. 144-55
- Chalmers, [Consciousness and its place in nature](#), section 6
- Chalmers, [Does conceivability entail possibility](#), appendix

5. Materialist responses

- Hill & McLaughlin, *from PPR symposium*
- Loar, *from PPR symposium*
- Yablo, [Concepts and consciousness](#)

6. Strong necessities and modal monism

- Chalmers, [Materialism and the metaphysics of modality](#), section 3
- Chalmers, [Does conceivability entail possibility](#), "From positive conceivability to possibility"

INDV 101: Philosophical Perspectives on the Individual

David Chalmers

E-mail: chalmers@arizona.edu

Phone: 621-7105

Web: <http://www.u.arizona.edu/~chalmers/>

Office hours: Thursday 1:30-3pm (Social Sciences 226A)

Class meeting (fall 1999): Monday/Wednesday 2-2:50pm, Modern Languages 350

Teaching Assistants:

- Allen Habib (anhabib@u.arizona.edu): Sections 6, 7, 13
- Ashley McDowell (ashleym@u.arizona.edu): Sections 4, 5, 10
- Seungbae Park (park@u.arizona.edu): Sections 3, 8, 11

Overview

This class is devoted to philosophical issues concerning the self and the mind. Some of the issues we will examine include:

- What is the self?
- What is the mind?
- What is the relationship between mind and body?
- Can consciousness be scientifically explained?
- How can we know about external reality?
- Do we have free will?
- Are animals conscious?
- Could computers have minds?

Texts

- *The Mind's I: Fantasies and Reflections on Self and Soul*, edited by Douglas Hofstadter and Daniel Dennett.
- *Doing Philosophy: An Introduction through Thought-Experiments*, by Theodore Schick Jr and Lewis Vaughn.

Further readings will be made available on this web page.

Assessment

There will be three short papers of 4-5 pages each, a midterm and a final exam, and occasional quizzes. Late papers will be penalized at a rate of one letter grade per day (except with medical documentation). No incompletes will be given, except for very good medical reasons. If you have special needs, let me know soon. Attendance at discussion section is required; if you miss more than two meetings, your grade will suffer.

[Second paper! \(due Wednesday December 1\)](#)

[First paper! \(due Monday October 4\)](#)

Tips on doing philosophy

Here are some very useful resources on how to write philosophy papers. Read these carefully!

- [Writing a Philosophy Paper](#) (Peter Horban)
- [Guide to the Study of Philosophy](#) (Garth Kemerling)
- [What Does He Want?](#) (Geoffrey Payzant)
- [Undergraduate Philosophical Writing](#) (Martin Young)
- [Writing a Philosophy Paper](#) (Sanford Levy)
- [Guidelines on Writing a Philosophy Paper](#) (James Pryor)

Schedule

Here is a very approximate week-by-week plan for the course, with associated readings. Note that this is very likely to be revised as things develop, and further readings will be added.

TOPIC 1. INTRODUCTION (1 week)

- DP, Chapter 1
- TMI, Introduction
- DP, pp. 215-217.

TOPIC 2: KNOWLEDGE AND SKEPTICISM (1 week)

- DP, Section 2.1
- Descartes, Meditations I & II (DP)
- DP, Section 7.1

TOPIC 3: DUALISM (2 weeks)

- DP, Section 2.1
- Beloff, [The mind-brain problem](#)
- Smullyan, An Unfortunate Dualist (TMI)

TOPIC 4: BEHAVIORISM (1 week)

- DP, Section 2.2
- Miedaner, The Soul of Martha, a beast (TMI)
- Miedaner, The Soul of the Mark III beast (TMI)
- [Behaviorism](#) (Internet Encyclopedia of Philosophy)

TOPIC 5: MATERIALISM (2 weeks)

- DP, Section 2.2
- Minsky, [Minds are simply what brains do](#)
- Dawkins, Selfish genes and selfish memes (TMI)
- Hofstadter, Prelude...ant fugue (TMI)

TOPIC 6: CONSCIOUSNESS (1-2 weeks)

- DP, Sections 2.4, 2.5
- Nagel, What is it like to be a bat? (TMI; also on the [web](#))
- Frank Jackson, [Epiphenomenal qualia](#)
- Chalmers, The puzzle of conscious experience (DP; also on the [web](#))

TOPIC 7: FREE WILL (2 weeks)

- DP, Chapter 3 (including readings)
- Smullyan, Is God a Taoist? (TMI)

TOPIC 8: PERSONAL IDENTITY (2 weeks)

- DP, Chapter 4 (including readings)
- Dennett, Where am I? (TMI)

TOPIC 9: ARTIFICIAL INTELLIGENCE (2-3 weeks)

- DP, Section 2.3

- Turing, Computing machinery and intelligence (TMI) (also on the [web](#))
- Searle, Minds, brains, and programs (TMI) (also on the [web](#))
- Bisson, [They're made of meat](#)

Web resources

Some useful resources on the web include:

- [Online papers on consciousness](#)
- [Consciousness in Philosophy: A Bibliography](#)
- [Consciousness in Science: A Bibliography](#)
- [Web resources related to consciousness, philosophy, and such](#)

Philosophy 80E: Paradoxes and Dilemmas

Compiled by [David Chalmers](#)

This is a web page for my course Philosophy 80E: Paradoxes and Dilemmas. It contains links to a number of papers and other resources related to paradoxes on the web. This page is still under construction and may be developed as time goes on.

Miscellaneous

- [Zeno's Coffeehouse](#) (Ron Barnette)
- [Contradictionary Web Page](#)

Logic/Language Paradoxes

- [Logic Puzzles](#)
- [Unexpected Hanging Paradox](#)
- [The Berry Paradox](#) (Gregory Chaitin)
- [The Paradox of the Question](#) (Ned Markosian)

Sorites Paradox & Vagueness

- [Dilemmas of Ambiguity and Vagueness](#) (Leon Felkins)
- [Understanding Vagueness](#) (Leon Felkins)
- [A Fuzzy Fairly Happy Face](#) (Patrick Grim)

Grue/Raven Paradoxes

- [Hempel's Ravens](#) (Franz Kiekeben)

Teletransporters, etc

- [Online papers on consciousness](#) (see especially [the self](#))

Time Travel

- [Time Travel](#) (Frank Kiekeben)

Prisoner's Dilemmas, etc

- [Paradoxes and Dilemmas](#) (Leon Felkins)
- [Social Dilemmas](#) (Leon Felkins) (plus [examples](#) and [games](#))
- [The Voter's Paradox](#) (Leon Felkins) (see also [here](#) and [here](#))
- [Prisoner's Dilemma](#) (Leon Felkins)
- [Prisoner's Dilemma](#) (Ethical Spectacle)
- [Prisoner's Dilemma](#) (Francis Heylighen)
- [Potential Solutions to the Prisoner's Dilemma](#) (Bjorn Brembs)
- [Interactive Prisoner's Dilemma](#)
- [The Tragedy of the Commons](#) (Garrett Hardin)

Newcomb's Problem

- [Newcomb's Paradox](#) (William Poundstone)
- [Newcomb's Paradox](#) (Franz Kiekeben)

Go to:

- [David Chalmers' home page](#)

From lan@U.Arizona.EDU Mon Sep 6 11:16:25 1999
Return-Path: <lan@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id LAA01867
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 6 Sep 1999 11:16:24 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFNR46HPQOB8SOD8@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 6 Sep 1999 18:17:21 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFNR44WPGWB8S95N@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 06 Sep 1999 18:17:19 -0700 (MST)
Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id LAA01863 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 06 Sep 1999 11:16:16 -0700
Received: from localhost (lan@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id SAA11696; Mon,
06 Sep 1999 18:17:15 -0700 (MST)
Date: Mon, 06 Sep 1999 18:17:14 -0700 (MST)
From: Lonnie A Nelson <lan@U.Arizona.EDU>
Subject: Re:a couple of questions...
In-reply-to: <Pine.HPX.4.10.9909011536350.14594-1000000@pavo.U.Arizona.EDU>
To: Logan T Trujillo <logant@U.Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909061801000.7299-1000000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

I have a couple of questions that have stuck in my mind, any and all input is welcome.

1) With regard to Denett's proposition that we "only think there is phenomenal consciousness" via many complex functions; my question, "is there "something it is like" to think that we have phenomenal consciousness? :)

2) With regard to phenomenally conscious states: It was given in the first lecture that a state "Y" is phenomenally conscious if there is something it is like to be in "Y". What followed was a list of "primary" or "foundationally" phenomenal conscious states. I may have mistaken this, but it seemed to imply to me that if a state was not "foundational" it did not belong on the list. This could have been because the list was of "primary" states, though that is not in my notes (i may have missed it). If this was meant to be a list of the kind of possible phenomenal states and not only of foundational or primary phenomenal states, it seems monumentally incomplete given that conscious experience is often multimodal. It would also seem to be the case that one could mix these primary states and come up with phenomenal experience that as a whole cannot be predicted accurately by the sum of its parts (a sort of systemic interaction of experience) as was proposed to be the case with "drunkenness" or "hypnotic states". Would it then be the case that one could think of the list that was put on the board as a "periodic table of phenomenal states" out of which all other complex states are constructed?

If so, what would one propose as the formula for "what it is like" to be in any given complex state (pick one) And / or how might we go about delineating empirically what the "formula" is for any given complex state?

Just thoughts,
-Lonnie

It is a common fate of all knowledge to begin as heresy and end as orthodoxy.

-Thomas Huxley

Lonnie A Nelson
Department of Psychology
Human Energy Systems Laboratory
University of Arizona
lan@u.arizona.edu

From chalmers@paradox.soc-sci.arizona.edu Mon Sep 6 16:45:51 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id QAA02562
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 6 Sep 1999 16:45:51 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JF02MMIB5SB8S9HW@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 6 Sep 1999 23:46:48 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTTP id <01JF02MKXII0B8T60J@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 06 Sep 1999 23:46:46 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id QAA02557 for scicon@paradox.soc-sci.arizona.edu; Mon,
06 Sep 1999 16:45:44 -0700
Date: Mon, 06 Sep 1999 16:45:44 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Re:a couple of questions...
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909062345.QAA02557@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: RO

Lonnie writes:

>2) With regard to phenomenally conscious states: It was given in the first
>lecture that a state "Y" is phenomenally conscious if there is something
>it is like to be in "Y". What followed was a list of "primary" or
>"foundationally" phenomenal conscious states. I may have mistaken this,
>but it seemed to imply to me that if a state was not "foundational" it did
>not belong on the list. This could have been because the list was of
>"primary" states, though that is not in my notes (i may have missed it).
>If this was meant to be a list of the kind of possible phenomenal states
>and not only of foundational or primary phenomenal states, it seems
>monumentally incomplete given that conscious experience is often
>multimodal. It would also seem to be the case that one could mix these
>primary states and come up with phenomenal experience that as a whole
>cannot be predicted accurately by the sum of its parts (a sort of systemic
>interaction of experience) as was proposed to be the case with
>"drunkenness" or "hypnotic states". Would it then be the case that one
>could think of the list that was put on the board as a "periodic table of
>phenomenal states" out of which all other complex states are constructed?

Well, the list wasn't intended to be taken too seriously, but there's
a sense in which it was trying to get at the "basic" sorts of
conscious states. As you say, there are any number of multimodal
states, but those get awfully complex. And one could argue that these
are decomposable to compositions of their various "modal" parts. If
they can't be so decomposed, then there may be something basic in
these states too (as we said e.g. for the category of "background
states"), and the only "periodic table" we would get would be an
indefinitely extendible one.

If it turns out that many or most such cases can be decomposed, though, then just perhaps it could turn out that we could have just a small number of classes of basic conscious states. In a way this might be a little bit like a periodic table. One big disanalogy, though, would be that within each class, there would be a huge number of "elements" (e.g. different sorts of sensory experiences) which may not be easily catalogable, and may not even be finite.

>If so, what would one propose as the formula for "what it is like" to
>be in any given complex state (pick one) And / or how might we go about
>delineating empirically what the "formula" is for any given complex state?

Good question. I suppose the simple answer to try would be to say that being in a composition of X and Y is like simultaneously being in X and being in Y. Of course, if decomposability fails, then your multimodal states won't really be compositions after all and this formula won't work. But I suspect that if decomposability fails, then there won't be any good formula, precisely because in that case the "whole" won't be a function of the "parts".

The question of how we go about determining whether a given complex state (e.g. hypnosis, drunkenness) is wholly decomposable into more specific elements is awfully tricky. I haven't thought about it a great deal and don't have any immediate ideas about it. It seems that phenomenology has to provide the primary database here. In some cases it seems clear phenomenologically that it would be possible to have states A and B without state C, so state C can't be reducible to A and B. (E.g., take C = the experience of humor and A and B = sensory experiences.) But in other cases it may be less clear. I think patient and reflective attention to phenomenology can go a long way here, but maybe there are "fringe" properties that are hard to resolve phenomenologically. It's not immediately obvious how empirical investigation can settle the matter, but maybe there are ways to bring it to bear. Any suggestions here are very welcome!

--Dave.

From lnielsen@azstarnet.com Wed Sep 8 13:07:15 1999

Return-Path: <lnielsen@azstarnet.com>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id NAA06698

for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 8 Sep 1999 13:07:14 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JFQ8VCH9V4B8T9LJ@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Wed, 8 Sep 1999 13:07:14 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU

(PMDF V5.1-12 #24137) with ESMTP id <01JFQ8V98RR4B8TB5Y@Telcom.Arizona.EDU>

for chalmers@Arizona.EDU; Wed, 08 Sep 1999 13:07:10 -0700 (MST)

Received: from cepheus.azstarnet.com (cepheus.azstarnet.com [169.197.56.195])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id NAA06694 for

<scicon@paradox.soc-sci.arizona.edu>; Wed, 08 Sep 1999 13:07:00 -0700

Received: from zippo (dialup19ip062.tus.azstarnet.com [169.197.39.62])

by cepheus.azstarnet.com (8.9.3/8.9.3) with SMTP id NAA09885 for

<scicon@paradox.soc-sci.arizona.edu>; Wed, 08 Sep 1999 13:06:09 -0700 (MST)

Date: Wed, 08 Sep 1999 13:11:33 -0700

From: Lis Nielsen <lnielsen@azstarnet.com>

Subject: RE: a couple of questions...

In-reply-to: <199909062345.QAA02557@paradox.soc-sci.arizona.edu>

To: scicon@paradox.soc-sci.arizona.edu

Reply-to: lnielsen@u.arizona.edu

Message-id: <NCBBJCIPCKAPOGDICDGPGEILCCAA.lnielsen@azstarnet.com>

MIME-version: 1.0
X-Mailer: Microsoft Outlook IMO, Build 9.0.2416 (9.0.2910.0)
Content-type: MULTIPART/MIXED; BOUNDARY="Boundary_(ID_Zj04cnl5dasI7crTdwNM5Q)"
Importance: Normal
X-Sent-via: StarNet <http://www.azstarnet.com/>
X-MSMail-Priority: Normal
X-MimeOLE: Produced By Microsoft MimeOLE V5.00.2314.1300
X-Priority: 3 (Normal)
Status: RO

This is a multi-part message in MIME format.

--Boundary_(ID_Zj04cnl5dasI7crTdwNM5Q)
Content-type: text/plain; charset=iso-8859-1
Content-transfer-encoding: 7bit

I have a couple of comments in response to Lonnie and Dave's chat about decomposability of complex conscious states, which went something like this:

> >If so, what would one propose as the formula for "what it is like" to
> >be in any given complex state (pick one) And / or how might we go about
> >delineating empirically what the "formula" is for any given
> complex state? (Lonnie)

> Good question. I suppose the simple answer to try would be to say
> that being in a composition of X and Y is like simultaneously being in
> X and being in Y. Of course, if decomposability fails, then your
> multimodal states won't really be compositions after all and this
> formula won't work. But I suspect that if decomposability fails, then
> there won't be any good formula, precisely because in that case the
> "whole" won't be a function of the "parts".
> ... It seems that
> phenomenology has to provide the primary database here. In some cases
> it seems clear phenomenologically that it would be possible to have
> states A and B without state C, so state C can't be reducible to A and
> B. (E.g., take C = the experience of humor and A and B = sensory
> experiences.) But in other cases it may be less clear. I think
> patient and reflective attention to phenomenology can go a long way
> here, but maybe there are "fringe" properties that are hard to resolve
> phenomenologically. It's not immediately obvious how empirical
> investigation can settle the matter, but maybe there are ways to bring
> it to bear. Any suggestions here are very welcome! (Dave)

My comments:

First, in our lab, we ask subjects to report on their experience of emotion (elicited by visual images) using an instrument called the Self-Assessment Manikin (Likert-like scales developed by Lang and colleagues at the University of Florida) that taps the dimensions of valence and arousal in emotional experience. (See attached document for a copy of the scales and instructions for their use). This procedure is based on the assumption that valence and arousal are two basic dimensions on which all emotional experiences vary. The support for this assumption comes from factor analysis work on emotional language, where valence and arousal turn out to be the two factors that best account for subjects' classifications of emotion words.

There are a number of advantages to this type of decomposition. It allows us to find reliable correlations between distinct features of phenomenal states and distinct components of emotional physiology. It allows us to compare groups of subjects on these distinct measures and determine whether

the two dimensions dissociate. For example, normal subjects will usually report higher arousal to images they have rated as highly negative or highly positive, while reporting lower arousal to neutral pictures. Some patients with frontal brain injuries give valence (positive or negative) ratings similar to normal subjects, but report diminished and undifferentiated arousal across image types. This suggests that the two components are separable.

The problem with the scales (and the dimensions in general) is that it is not at all clear (though I tend to vacillate on this issue) that they represent more than linguistic classifications subjects are able to make. Moreover, the scales are presented as linear scales with specific endpoints. Negative and positive are presented on the same dimension with neutral as a midpoint, ruling out the possibility of experiences being a little negative and a little positive. The advantages of simplicity for data collection can lead to serious limitations in our understanding of the fundamental dimensions (or elements) of conscious experience. That said, it does seem possible for us to abstract dimensions from our emotional experience, just as we can extract visual dimensions from our experience of objects (color, shape, etc.). Most subjects have no problem making the ratings, and report that they were able to rate their experiences on both dimensions.

For a science of phenomenal experience to develop, one needs to begin by assuming some sorts of dimensions, in order to make progress. The problem comes in making assumptions about what is basic. Reservations that we have had about what counts as basic in emotional experience have led us to consider the use of open-ended emotion narratives as data. This demands that we develop a coding scheme that can render these narratives scientifically interpretable. In the development of this scheme, we begin by making some general assumptions about the elements that constitute emotional experiences (based on a lot of background theory) without committing ourselves to any particular map of the phenomenal space. We remain open to the question of what is basic and what is complex, hoping that the data will reveal this. In implementing this type of measurement in our lab, we will also be using the SAM scales, to determine whether there are any reliable mappings between the classifications our data reveal and the SAM dimensions.

My guess is that we may be able to identify certain emotion prototypes, and that these may turn out to vary along certain basic dimensions. Perhaps these will look something like the SAM dimensions, but I expect they will be a bit more complex. I do not think we will be able to identify primitive or basic qualia into which emotional experiences can be reduced. I am always surprised at the desire to implement some form of atomic reductionism at the phenomenal level, and wonder whether it doesn't result from our entrenchment in a predominantly reductive-materialist Western scientific mindset.

Lis

Lis Nielsen
Department of Psychology
University of Arizona
Tucson, AZ 85721-0068
USA

lnielsen@u.arizona.edu

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 8 19:06:38 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id TAA07541

for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 8 Sep 1999 19:06:38 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JFQLF17J00B8TABQ@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Wed, 8 Sep 1999 19:06:41 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU

(PMDF V5.1-12 #24137) with ESMTP id <01JFQLF07EZKB8T2JA@Telcom.Arizona.EDU>

for chalmers@Arizona.EDU; Wed, 08 Sep 1999 19:06:40 -0700 (MST)

Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu

(8.9.3/8.9.3) id TAA07518 for scicon; Wed, 08 Sep 1999 19:06:34 -0700

Date: Wed, 08 Sep 1999 19:06:34 -0700

From: David Chalmers <chalmers@arizona.edu>

Subject: physics, emotion, and baseball

To: scicon@paradox.soc-sci.arizona.edu

Message-id: <199909090206.TAA07518@paradox.soc-sci.arizona.edu>

MIME-version: 1.0

Content-type: TEXT/PLAIN; CHARSET=US-ASCII

Status: R

Some miscellaneous comments on things that have come up:

(1) Logan raises the interesting possibility that as well as the role that intrinsic properties play in grounding extrinsic properties (on a panprotopsychoist view), there may also be the possibility of further "direct interactions" between intrinsic properties which would then be empirically testable. I guess my question here is what the principled difference is between this and standard "extrinsic" interactions. After all, when one particle affects another in standard ways, a panprotopsychoist can construe this as the intrinsic properties of one directly having effects on the intrinsic properties of the other.

Of course physics doesn't "see" the intrinsic part and characterizes only the extrinsic structure. But the same would go for Logan's new "direct" interactions, it seems to me. Even in parapsychology, one could characterize the causal structure extrinsically. All these would have an extrinsic structure that an objective science could characterize, and all would basic intrinsic properties that are really doing the causing and being affected. On either case, the extrinsic structure is just a way of getting at the basic causal relations among the intrinsic properties. So I'm not sure that I see the difference in kind here.

(2) Lis raises lots of interesting issues re her study of dimensions of emotional experience. The distinction between identifying dimensions of experience and reducing experience to basic components is important. It may well be that we can do the former but not the latter. For certain experiences, we can do both: e.g. for color, it's arguable that the experience is reducible to the components of hue, brightness, and saturation. But maybe it will turn out that for emotional experience, we will find at best a limited number of parameters that don't come close to exhausting the phenomenal character of the experience. I guess it's partly an empirical question whether the sort of reductive analysis that works for color will work elsewhere, but there's no obvious reason to believe that it should.

Philosophically, one might put the issue here in terms of supervenience. Just say one has identified a number of parameters of experience, and are wondering whether an experience is reducible to those parameters. I guess this comes down to whether the character of the experience supervenes on those parameters. I.e., would it be possible in principle (even logically possible) to have two different

experiences with the same values of those parameters but different overall character? With color, one might well argue that it is logically impossible to have two experiences with the same hue, saturation, and brightness, but a different color quality. (Though maybe there are issues about further parameters such as shimmer, etc.) But with emotion, for any set of parameters we have identified at least so far, the parameters don't seem to determine the quality of the experience in the same way. Maybe that just means we have to come up with better parameters, but again, it's not obvious that any full set of such parameters should exist.

I note that when I put forward the (very tentative and speculative) suggestion about five or six basic sorts of phenomenal states, I wasn't really suggesting anything quite this reductive. Here, the basic categories were as broad as perceptual experiences, emotional experiences, cognitive experiences, etc, and the claim was in effect that the overall character of a subject's experiences is fully determined by the character of these components. This is consistent with the view that within each category, there may be a huge variety of states that are not reducible to any more basic building blocks (as e.g. Lis suggests re emotion).

In effect we have two questions: (i) whether a subject's overall state of experience is determined by the character of their experiences in the five or six categories, and (ii) whether the character of an experiences within those five or six categories is determined by some set of more basic parameters. Both of these seem to be interesting open questions, and the answer isn't obvious (at least to me) one way or another.

(3) Re Lonnie's and Joel's discussion of the baseball player: Actually there is recent empirical work that bears on this. The research of Milner and Goodale, which we'll be discussing, suggests that there are two pathways in the visual system, one for making cognitive and conscious identifications, and one for controlling online motor action. In certain cases (e.g. due to brain damage), representations in these two pathways can come apart, and one finds subjects who can perform fine-grained motor actions on objects even when they apparently don't have conscious knowledge of those objects. (E.g., a subject who can "post" a letter through a horizontal or vertical slot even though she doesn't consciously know whether it is horizontal or vertical.) It even turns out that one can find such subtle versions of such dissociations in ordinary subjects under certain conditions, and in these cases one finds that how a subject acts with respect to (say) the position of an object may be out of line with their conscious judgments.

It's not out of the question that Joel's observations re the phenomenology of baseball could be partly grounded in this framework. It could turn out that representations in the "cognitive" system are less fine-grained than representations in the "motor" system. Milner and Goodale believe that the contents of visual consciousness are determined by the cognitive system, not the motor system. If that's the case, then one would expect that one would be able to performed fine-grained motor tasks (e.g. catching a baseball) that rely on visual information that is not present in consciousness at all).

--Dave.

From logant@U.Arizona.EDU Wed Sep 1 15:50:47 1999
Return-Path: <logant@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id PAA31917
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 1 Sep 1999 15:50:47 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFGMKXE2U0B8SC3K@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 1 Sep 1999 15:51:45 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFGMKVZOV4B8S2L5@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 01 Sep 1999 15:51:44 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id PAA31912 for
<scicon@paradox.soc-sci.arizona.edu>; Wed, 01 Sep 1999 15:50:43 -0700
Received: from localhost (logant@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id PAA23539 for
<scicon@paradox.soc-sci.arizona.edu>; Wed, 01 Sep 1999 15:51:42 -0700 (MST)
Date: Wed, 01 Sep 1999 15:51:42 -0700 (MST)
From: Logan T Trujillo <logant@U.Arizona.EDU>
Subject: Functional role of intrinsic consciousness
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909011536350.14594-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

This question is directed at David, but may be replied by anyone in the class who has something to say. I noticed that when David presented his non-reductive strategy to investigating consciousness, he drew a single one-way arrow pointing from physical processes to consciousness. David, were you trying to imply that the causal-functional relationship between physical and qualitative processes is only one way (from physical to consciousness)? What reasons do we have to think that it couldn't go both ways (kind of like J. Sarfatti's idea of backaction between mental and physical)? I ask this because if we want to introduce consciousness as an explanatory primitive, and if we further wish to characterize it as playing a causal-functional role in the universe, then it may be that our current account of non-conscious physical processes is incomplete in that the intrinsic nature to material particles may be playing a role in the dynamics of physical systems that are currently characterized in terms of extrinsic relationships. Such a possibility could produce empirically testable consequences.

Such a possibility would be consistent with both pan-experiential and interactionist viewpoints in that the ontology could be construed in the sense of the former - consciousness (or at least qualitative experience) is an intrinsic property to material processes - yet the dynamics could be understood in the sense of the latter in that there could be a functional, interactive relationship between the intrinsic and extrinsic parameters of physical reality. Physical reality would still be causally closed, but only if one takes both its inner and outer aspects into account.

Just a thought.

Best,

Logan Trujillo
Psychology

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 1 20:31:55 1999
Return-Path: <chalmers@arizona.edu>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id UAA32244
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 1 Sep 1999 20:31:55 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
id <01JFGWEF055CB8SBV9@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 1 Sep 1999 20:32:51 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDf V5.1-12 #24137) with ESMTP id <01JFGWEDQ3EOB8S7FC@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 01 Sep 1999 20:32:49 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id UAA32236 for scicon@paradox.soc-sci.arizona.edu; Wed,
01 Sep 1999 20:31:45 -0700
Date: Wed, 01 Sep 1999 20:31:45 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Re: Functional role of intrinsic consciousness
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909020331.UAA32236@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

Logan writes:

>This question is directed at David, but may be replied by anyone in the
>class who has something to say. I noticed that when David presented his
>non-reductive strategy to investigating consciousness, he drew a single
>one-way arrow pointing from physical processes to consciousness. David,
>were you trying to imply that the causal-functional relationship between
>physical and qualitative processes is only one way (from physical to
>consciousness)? What reasons do we have to think that it couldn't go both
>ways (kind of like J. Sarfatti's idea of backaction between mental and
>physical)?

Actually, I didn't mean to rule out that the relation could be
two-way. In my alphabetical taxonomy, the one-way relationship is a
type-E dualism (epiphenomenalism), and the two-way relationship is a
type-D dualism (D for Descartes, i.e. Cartesian dualism, or
interactionism).

The reason many people are suspicious of the type-D view is that it
seems to introduce gaps in physical causation; i.e., it suggests that
the physical world is not "causally closed". And many people think
that the evidence from physics is that physics is causally closed, and
that there is no room for "spooky causation". Of course some people
deny that, and suggest that in particular quantum mechanics may leave
room for interaction between physical and mental.

Many are suspicious of the type-E view for a different reason, namely
that it suggests that consciousness has no effect on the physical
world. And seems very counterintuitive to hold that a pain doesn't
cause our pain reaction, that the visual experience of the sunset
doesn't cause me to keep looking at it, that my consciousness plays no
causal role in my writing books about consciousness, etc. But then
one might respond that the epiphenomenalist can explain the
counterintuitiveness at least to some extent, by noting that there are
strong regularities between experiences and consequent physical
events, so that we tend to infer a causal connection even when it is
not there.

Of course all that is controversial. My own view is that if arguments
against materialism are accepted, then the conclusion ought to be a

disjunction of (i.e. a choice between) interactionism, epiphenomenalism, and panprotopsydism (i.e. the type-D, type-E, and type-F views I mentioned). The choice between those depends on a lot of subtle matters and is by no means obvious, and may to some extent depend on where a science of consciousness ends up going.

>I ask this because if we want to introduce consciousness as an
>explanatory primitive, and if we further wish to characterize it as playing a
>causal-functional role in the universe, then it may be that our current
>account of non-conscious physical processes is incomplete in that the
>intrinsic nature to material particles may be playing a role in the
>dynamics of physical systems that are currently characterized in terms of
>extrinsic relationships. Such a possibility could produce empirically
>testable consequences.

> Such a possibility would be consistent with both pan-experiential
>and interactionist viewpoints in that the ontology could be construed in
>the sense of the former - consciousness (or at least qualitative
>experience) is an intrinsic property to material processes - yet the
>dynamics could be understood in the sense of the latter in that there
>could be a functional, interactive relationship between the intrinsic and
>extrinsic parameters of physical reality. Physical reality would still be
>causally closed, but only if one takes both its inner and outer aspects into
>account.

Hmm, your talk of the "intrinsic aspects of material reality" suggests the panprotopsydism view to me. Of course on that view, the intrinsic aspects have causal effects (they are the ultimate causes of all physical action!), but on this view, the physical world will still be causally closed, so it isn't a type-D view. I'm not sure just how you mean the view to have empirical consequences, though. On my reading of such a view, the intrinsic aspects play a causal role in all physical interactions (they are the ultimate things doing the causing), so it's not as if there will be a new and special set of interactions that they do the causing in. So it's not clear that the view makes empirical predictions (at least about third-person matters such as the motion of particles) that are distinct from those made by ordinary physical theory.

If, on the other hand, you want to introduce phenomenal properties as a new set of properties at the bottom level without having them underlie all causal interactions, then it seems that you'll be regarding them as a further property like mass, etc, right at the bottom. Then it's true that you could have new laws and new interactions governing them. But now you'll be up against the problem that there is no apparent evidence for them in physics, and that the view will thus require revolutionizing our well-supported physical theories to incorporate consciousness. Maybe this could happen, but it's going out on a big limb. Basically, on this view, we'll say that there's a certain physical parameter C (to be discovered) such that C has consciousness as its intrinsic nature. Maybe it would be easier to say that some existing physical parameter (mass or spin, say) has consciousness as its intrinsic nature, and that it interacts with all the others. Then, it seems to me, you might get the benefits of the view without the costs.

Feel free to clarify...

--Dave.

From logant@U.Arizona.EDU Thu Sep 2 21:02:38 1999

Return-Path: <logant@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA01587

for <chalmers@paradox.soc-sci.arizona.edu>; Thu, 2 Sep 1999 21:02:38 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JFIBQYVP20B8SPJA@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Thu, 2 Sep 1999 21:03:38 MST

Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

with ESMTP id <01JFIBQXQXSGB8S5L9@Telcom.Arizona.EDU> for

chalmers@Arizona.EDU; Thu, 02 Sep 1999 21:03:37 -0700 (MST)

Received: from localhost (logant@localhost)

by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id VAA00859; Thu,

02 Sep 1999 21:03:36 -0700 (MST)

Date: Thu, 02 Sep 1999 21:03:36 -0700 (MST)

From: Logan T Trujillo <logant@U.Arizona.EDU>

Subject: Re: Functional role of intrinsic consciousness

In-reply-to: <199909020331.UAA32236@paradox.soc-sci.arizona.edu>

To: David Chalmers <chalmers@Arizona.EDU>

Cc: scicon@paradox.soc-sci.arizona.edu

Message-id: <Pine.HPX.4.10.9909021939460.4890-100000@orion.U.Arizona.EDU>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: RO

On Wed, 1 Sep 1999, David Chalmers wrote:

> My own view is that if arguments
> against materialism are accepted, then the conclusion ought to be a
> disjunction of (i.e. a choice between) interactionism,
> epiphenomenalism, and panprotopsychism (i.e. the type-D, type-E, and
> type-F views I mentioned). The choice between those depends on a lot
> of subtle matters and is by no means obvious, and may to some extent
> depend on where a science of consciousness ends up going.
>

[Logan]

> >I ask this because if we want to introduce consciousness as an
> >explanatory primitive, and if we further wish to characterize it as playing a
> >causal-functional role in the universe, then it may be that our current
> >account of non-conscious physical processes is incomplete in that the
> >intrinsic nature to material particles may be playing a role in the
> >dynamics of physical systems that are currently characterized in terms of
> >extrinsic relationships. Such a possibility could produce empirically
> >testable consequences.

> > Such a possibility would be consistent with both pan-experiential
> >and interactionist viewpoints in that the ontology could be construed in
> >the sense of the former - consciousness (or at least qualitative
> >experience) is an intrinsic property to material processes - yet the
> >dynamics could be understood in the sense of the latter in that there
> >could be a functional, interactive relationship between the intrinsic and
> >extrinsic parameters of physical reality. Physical reality would still be
> >causally closed, but only if one takes both its inner and outer aspects into
> >account.

[David]

> Hmm, your talk of the "intrinsic aspects of material reality" suggests
> the panprotopsychist view to me. Of course on that view, the
> intrinsic aspects have causal effects (they are the ultimate causes of
> all physical action!), but on this view, the physical world will still
> be causally closed, so it isn't a type-D view.

[Logan]

I think that we should clarify what we mean by "physical". On a

panprotopsychist account, the intrinsic aspects to material particles would still be "physical". However they have a distinction from physical properties that are characterized extrinsically (mass, charge, etc).

It is this distinction that proponents of this view seek to utilize in an account for the "difference in kind" of experiential properties from the extrinsic properties we know. Another terminology that could be used is "material/immaterial" in place of "extrinsic/intrinsic". "Material" properties may be conceived as those arising due to extrinsic relationships, "immaterial" properties are those due to intrinsic.

Now on a panprotopsychist viewpoint the intrinsic properties certainly may be construed as carrying the causal weight of the extrinsic interactions (see below). However physics for the most part maps out extrinsic properties (an argument against this may be made in the case of spin; spin is considered an "intrinsic" property because it does not depend explicitly on spatiotemporal parameters. However it does have extrinsic effects and thus may be characterized in terms of this relational ability). The intrinsic nature to reality may account for this causally, but the possibility remains that certain physical processes exist that are purely "immaterial", i.e. interact via direct relationships among intrinsic properties. These may have influences extrinsically in the sense that these interactions produce changes in the intrinsic state space of a system in a manner that carries over into the portion of the intrinsic causal network underlying extrinsic properties. Although panprotopsychist, this view is also interactionist in the sense of a material/immaterial categorization to the ONE physical reality that is causally closed on a global (intrinsic/extrinsic) scale (this would be a conjunction of type D and F).

I guess it comes down to the question, can the possible set of intrinsic property relations and interactions outrun the possible set of extrinsic property relations and interactions? I see no reason why this couldn't be the case, and recent arguments based in physics approaches to understanding consciousness, as well as the data from parapsychology seem to suggest that this might be the case (although this is very controversial). An analogy from mathematics might help to illustrate: by definition, all real numbers are complex numbers, but not all complex numbers are real numbers. I would argue in analogy that the intrinsic stands to the complex numbers as the extrinsic stands to the reals. To be extrinsic is to have an accompanying intrinsic component; but being intrinsic does not imply that one has an accompanying component in any direct sense.

[David]

I'm not sure just how

> you mean the view to have empirical consequences, though.

On my

> reading of such a view, the intrinsic aspects play a causal role in
> all physical interactions (they are the ultimate things doing the
> causing), so it's not as if there will be a new and special set of
> interactions that they do the causing in. So it's not clear that the
> view makes empirical predictions (at least about third-person matters
> such as the motion of particles) that are distinct from those made by
> ordinary physical theory.

The consequences may be as everyday as a solution to phenomenal binding, to the existence of a freewill, to the more controversial "anomalies of consciousness" data mentioned above. These are all phenomena that seem to defy extrinsic relational explanations.

>

> If, on the other hand, you want to introduce phenomenal properties as
> a new set of properties at the bottom level without having them
> underlie all causal interactions, then it seems that you'll be
> regarding them as a further property like mass, etc, right at the
> bottom. Then it's true that you could have new laws and new
> interactions governing them. But now you'll be up against the problem
> that there is no apparent evidence for them in physics, and that the
> view will thus require revolutionizing our well-supported physical
> theories to incorporate consciousness. Maybe this could happen, but
> it's going out on a big limb. Basically, on this view, we'll say that
> there's a certain physical parameter C (to be discovered) such that C
> has consciousness as its intrinsic nature. Maybe it would be easier
> to say that some existing physical parameter (mass or spin, say) has
> consciousness as its intrinsic nature, and that it interacts with all
> the others. Then, it seems to me, you might get the benefits of the
> view without the costs.
>

[Logan]

This almost sounds like my viewpoint except that I suspect that phenomenal consciousness does underlie all extrinsic interactions. It is just that phenomenal properties may interact in a manner over and above the purely extrinsic interactions. Again I don't think there is any good reason to deny this, especially in the face of the fact that physics, when first conceived, split off primary qualities from secondary qualities a priori, with the former deemed proper subject of study. The whole edifice of physics is rooted in this distinction. Thus when trying to slip secondary (phenomenal) properties "under the carpet", so to speak, what a priori justification do we have to maintain that the relationships within and between intrinsics/extrinsic properties must match up in a one-to-one manner?

> Feel free to clarify...

>
> --Dave.
>

Thanks for the opportunity!

Logan

From press@U.Arizona.EDU Mon Sep 6 12:11:20 1999

Return-Path: <press@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id MAA02115

for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 6 Sep 1999 12:11:20 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JFNT1ALDKGB8SNV0@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Mon, 6 Sep 1999 19:12:18 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU

(PMDF V5.1-12 #24137) with ESMTP id <01JFNT19D45SB8SSEM@Telcom.Arizona.EDU>

for chalmers@Arizona.EDU; Mon, 06 Sep 1999 19:12:16 -0700 (MST)

Received: from f1n8.u.arizona.edu

(IDENT:press@f1n8.U.Arizona.EDU [128.196.137.108])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id MAA02111 for

<scicon@paradox.soc-sci.arizona.edu>; Mon, 06 Sep 1999 12:11:14 -0700

Received: from localhost (press@localhost) by f1n8.u.arizona.edu (8.8.8/8.8.8)

with ESMTP id TAA02160 for <scicon@paradox.soc-sci.arizona.edu>; Mon,

06 Sep 1999 19:12:12 -0700

Date: Mon, 06 Sep 1999 19:12:12 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>
Subject: Dave's Principle of Structural Coherence
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909061903360.25050-100000@f1n8.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R0

All-

This post deals primarily with Daves "Facing Up to the Problem of Consciousness," so I suppose it is directed somewhat at him (its terribly useful to have the author of the paper so readily available.) However, I think that the issue that follows may lend itself to lots of group input, at least for those interested. The point Im going to try to make is at least largely an empirical one, but I am pretty certain that my own knowledge of this subject only scratches the surface of the available evidence. So anyone who can think of other experimental results or examples that either confirm or disconfirm my point would be welcome to add their evidence to the pile.

What I want to write about here is Daves Principle of Structural Coherence, which is intended as the backbone of his scientific theory of consciousness. But before I do, I should say that I think this part of his paper (and book) is tremendously important. This is because it defends him against one of the stronger arguments against dualism, namely the argument that dualism amounts to just giving up on explaining consciousness. Many materialists refuse to even consider dualism because most dualists claim that consciousness must forever remain a mystery. (See Dennetts Consciousness Explained, p. 37 for an example of this.) However, if something like Daves Principle of Structural Coherence could be confirmed, this might take the sting out of dualism by opening phenomenal consciousness to scientific investigation.

The problem is that several psychological facts seem to disconfirm the Principle of Structural Coherence as outlined in section seven of "Facing Up to the Problem of Consciousness." The basic idea here is that the structure of "consciousness" (phenomenal consciousness) and the structure of "awareness" (direct availability for global control) mirror each other. The following are examples where the two seem (at least to me) to diverge:

Awareness without Consciousness - In the Nelkin paper, "What is Consciousness," he defends several examples of what he takes to be intentionality without introspection, which I take to be awfully similar to awareness without consciousness. If a persons brain can be shown to be carrying on complex and integrated intentional behavior while they themselves report no consciousness of these facts, it seems that there is information in their brain that is available for global control but not conscious.

Unified Consciousness vs. Dis-unified Awareness - In the Bisiach paper, "The (Haunted) Brain and Consciousness," he starts the section on "The modularity of consciousness" with the claim that "Data from investigation of brain-damaged subjects show dramatically how C2 is far from being a unitary process (and how vain it would be to try to map C1 onto C2)." If awareness is generally dis-unified, with the brain processes responsible for various tasks working largely independently, at different speeds, and so on, this does not seem to correspond to our phenomenological consciousness of the goings on in our mind.

Detailed Awareness vs. Approximate Consciousness - This example is just based on personal observation, though I wouldn't be surprised to find that more careful experiments have been done. Many tasks seem to require more detailed awareness than we are conscious of when we perform them. For example, if you ask a baseball player to guess the angle and initial velocity of the fly ball coming off the bat, even the most mathematically and scientifically astute baseball player will only be able to roughly estimate these quantities. The same player, however, will successfully position himself or herself to catch the ball. It seems that the player needs to be aware of data with a significantly smaller margin of error than he or she would report being conscious of.

There are lots of responses that Dave can make to these sorts of observations. Indeed, as I am sure that other people have raised similar questions, I fully expect that he has ready responses for most of them. I realize that I've been fairly vague about giving specific examples of the sorts of phenomena I think conflict with the Structural Coherence Principle, but I wanted to get the main points across. Perhaps if Dave doesn't have a way of dismissing the lot in one fell swoop, we'll get into talking about specific cases. If anyone has examples of phenomena that seem to fit the bill, let me know.

Thanks,

Joel

PS - I really do know how to use apostrophes. MS Word and PINE don't seem to get along in this respect.

From lan@U.Arizona.EDU Tue Sep 7 05:36:34 1999
Return-Path: <lan@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id FAA03912
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 7 Sep 1999 05:36:34 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFOTJ83E20B8TGEU@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 7 Sep 1999 12:37:33 MST
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU)
by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JFOTJ6W4CWB8T40V@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Tue, 07 Sep 1999 12:37:31 -0700 (MST)
Received: from localhost (lan@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id MAA20793; Tue,
07 Sep 1999 12:37:30 -0700 (MST)
Date: Tue, 07 Sep 1999 12:37:30 -0700 (MST)
From: Lonnie A Nelson <lan@U.Arizona.EDU>
Subject: Re: Dave's Principle of Structural Coherence
In-reply-to: <199909062333.QAA02540@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: press@U.Arizona.EDU, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909071231590.10941-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

> apparently not present in awareness.

>

> >Detailed Awareness vs. Approximate Consciousness - This example is
> >just based on personal observation, though I wouldn't be surprised to
> >find that more careful experiments have been done. Many tasks seem
> >to require more detailed awareness than we are conscious of when we

> >perform them. For example, if you ask a baseball player to guess the
> >angle and initial velocity of the fly ball coming off the bat, even
> >the most mathematically and scientifically astute baseball player
> >will only be able to roughly estimate these quantities. The same
> >player, however, will successfully position himself or herself to
> >catch the ball. It seems that the player needs to be aware of data
> >with a significantly smaller margin of error than he or she would
> >report being conscious of.
>
> I think I would say that in these cases the player isn't aware of the
> information, either. As you note yourself, the information isn't
> available for verbal report. In effect it is playing a local control
> role in the regulation of certain sorts of specific and practiced
> behavior, but it isn't available in the sort of general and global
> manner required for awareness. So again, if awareness is delineated
> appropriately, there's no dissociation.

It seems to me that the information that is being referred to here is of a cognitive domain (mathematically astute athlete) would be dealing in arbitrary units with which to quantify his (or her) experience. The player would not experience the ball in terms of degrees, as a "degree" is not a naturally occurring concept to most of us, we give estimates of these sorts of measures in order to communicate, but when you are catching a fly ball, you are accessing a "natural" schema of prior experience. "What this particular fly ball is like" is very apparent to the player, and if you could be shown his (or her) mind, from the inside they could direct you to which set of memories merged to tell them how to behave. I realize that this came off sounding rather self assured, but really the above is only a string of hypotheses generated by my own experience. And therefore are entirely open to alternative interpretation...Anyone?

--Lonnie

> --Dave. >

It is a common fate of all knowledge to begin as heresy and end as orthodoxy.

-Thomas Huxley

Lonnie A Nelson
Department of Psychology
Human Energy Systems Laboratory
University of Arizona
lan@u.arizona.edu

From press@U.Arizona.EDU Wed Sep 8 17:15:23 1999
Return-Path: <press@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA07282
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 8 Sep 1999 17:15:23 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFQHJ3L5J4B8TMC3@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 8 Sep 1999 17:15:26 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFQHJ1V074B8TBS0@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 08 Sep 1999 17:15:24 -0700 (MST)
Received: from f1n8.u.arizona.edu
(IDENT:press@f1n8.U.Arizona.EDU [128.196.137.108])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA07274 for
<scicon@paradox.soc-sci.arizona.edu>; Wed, 08 Sep 1999 17:15:17 -0700

Received: from localhost (press@localhost) by f1n8.u.arizona.edu (8.8.8/8.8.8) with ESMTP id RAA36722; Wed, 08 Sep 1999 17:15:20 -0700
Date: Wed, 08 Sep 1999 17:15:19 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>
Subject: Re: Dave's Principle of Structural Coherence
In-reply-to: <Pine.HPX.4.10.9909071231590.10941-100000@pavo.U.Arizona.EDU>
To: Lonnie A Nelson <lan@U.Arizona.EDU>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909081656420.30220-100000@f1n8.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

Lonnie,

Actually, misleading remarks about mathematics to the side, I was thinking of the baseball player's visual representation. When I catch a fly ball, it seems to me as though the information contained in my phenomenal visual field is less precise than that needed to catch the ball. I'm not sure about this either. For one thing, I haven't caught many fly balls lately, so maybe I'm remembering it wrong. Or maybe I've just discovered the reason that I'm only a mediocre outfielder.

Thanks for the reply,

Joel

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 8 19:06:38 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217]) by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id TAA07541 for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 8 Sep 1999 19:06:38 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137) id <01JFQLF17J00B8TABQQ@Telcom.Arizona.EDU> for chalmers@paradox.soc-sci.arizona.edu; Wed, 8 Sep 1999 19:06:41 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137) with ESMTP id <01JFQLF07EZKB8T2JA@Telcom.Arizona.EDU> for chalmers@Arizona.EDU; Wed, 08 Sep 1999 19:06:40 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) id TAA07518 for scicon; Wed, 08 Sep 1999 19:06:34 -0700
Date: Wed, 08 Sep 1999 19:06:34 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: physics, emotion, and baseball
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909090206.TAA07518@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

Some miscellaneous comments on things that have come up:

(1) Logan raises the interesting possibility that as well as the role that intrinsic properties play in grounding extrinsic properties (on a panprotopsychnist view), there may also be the possibility of further "direct interactions" between intrinsic properties which would then be empirically testable. I guess my question here is what the principled difference is between this and standard "extrinsic" interactions. After all, when one particle affects another in standard ways, a panprotopsychnist can construe this as the intrinsic properties of one directly having effects on the intrinsic properties of the other.

Of course physics doesn't "see" the intrinsic part and characterizes

only the extrinsic structure. But the same would go for Logan's new "direct" interactions, it seems to me. Even in parapsychology, one could characterize the causal structure extrinsically. All these would have an extrinsic structure that an objective science could characterize, and all would basic intrinsic properties that are really doing the causing and being affected. On either case, the extrinsic structure is just a way of getting at the basic causal relations among the intrinsic properties. So I'm not sure that I see the difference in kind here.

(2) Lis raises lots of interesting issues re her study of dimensions of emotional experience. The distinction between identifying dimensions of experience and reducing experience to basic components is important. It may well be that we can do the former but not the latter. For certain experiences, we can do both: e.g. for color, it's arguable that the experience is reducible to the components of hue, brightness, and saturation. But maybe it will turn out that for emotional experience, we will find at best a limited number of parameters that don't come close to exhausting the phenomenal character of the experience. I guess it's partly an empirical question whether the sort of reductive analysis that works for color will work elsewhere, but there's no obvious reason to believe that it should.

Philosophically, one might put the issue here in terms of supervenience. Just say one has identified a number of parameters of experience, and are wondering whether an experience is reducible to those parameters. I guess this comes down to whether the character of the experience supervenes on those parameters. I.e., would it be possible in principle (even logically possible) to have two different experiences with the same values of those parameters but different overall character? With color, one might well argue that it is logically impossible to have two experiences with the same hue, saturation, and brightness, but a different color quality. (Though maybe there are issues about further parameters such as shimmer, etc.) But with emotion, for any set of parameters we have identified at least so far, the parameters don't seem to determine the quality of the experience in the same way. Maybe that just means we have to come up with better parameters, but again, it's not obvious that any full set of such parameters should exist.

I note that when I put forward the (very tentative and speculative) suggestion about five or six basic sorts of phenomenal states, I wasn't really suggesting anything quite this reductive. Here, the basic categories were as broad as perceptual experiences, emotional experiences, cognitive experiences, etc, and the claim was in effect that the overall character of a subject's experiences is fully determined by the character of these components. This is consistent with the view that within each category, there may be a huge variety of states that are not reducible to any more basic building blocks (as e.g. Lis suggests re emotion).

In effect we have two questions: (i) whether a subject's overall state of experience is determined by the character of their experiences in the five or six categories, and (ii) whether the character of an experiences within those five or six categories is determined by some set of more basic parameters. Both of these seem to be interesting open questions, and the answer isn't obvious (at least to me) one way or another.

(3) Re Lonnie's and Joel's discussion of the baseball player: Actually

there is recent empirical work that bears on this. The research of Milner and Goodale, which we'll be discussing, suggests that there are two pathways in the visual system, one for making cognitive and conscious identifications, and one for controlling online motor action. In certain cases (e.g. due to brain damage), representations in these two pathways can come apart, and one finds subjects who can perform fine-grained motor actions on objects even when they apparently don't have conscious knowledge of those objects. (E.g., a subject who can "post" a letter through a horizontal or vertical slot even though she doesn't consciously know whether it is horizontal or vertical.) It even turns out that one can find such subtle versions of such dissociations in ordinary subjects under certain conditions, and in these cases one finds that how a subject acts with respect to (say) the position of an object may be out of line with their conscious judgments.

It's not out of the question that Joel's observations re the phenomenology of baseball could be partly grounded in this framework. It could turn out that representations in the "cognitive" system are less fine-grained than representations in the "motor" system. Milner and Goodale believe that the contents of visual consciousness are determined by the cognitive system, not the motor system. If that's the case, then one would expect that one would be able to performed fine-grained motor tasks (e.g. catching a baseball) that rely on visual information that is not present in consciousness at all).

--Dave.

From logant@U.Arizona.EDU Sat Sep 11 17:11:01 1999
Return-Path: <logant@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA03034
for <chalmers@paradox.soc-sci.arizona.edu>; Sat, 11 Sep 1999 17:11:00 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFU09NXFB4B8TT37@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sat, 11 Sep 1999 17:11:02 MST
Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JFU09MHIOWB80JYA@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Sat, 11 Sep 1999 17:11:00 -0700 (MST)
Received: from localhost (logant@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id RAA24092; Sat,
11 Sep 1999 17:10:59 -0700 (MST)
Date: Sat, 11 Sep 1999 17:10:59 -0700 (MST)
From: Logan T Trujillo <logant@U.Arizona.EDU>
Subject: Re: physics, emotion, and baseball
In-reply-to: <199909090206.TAA07518@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909111549300.6567-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

On Wed, 8 Sep 1999, David Chalmers wrote:

> Some miscellaneous comments on things that have come up:
>
> (1) Logan raises the interesting possibility that as well as the role
> that intrinsic properties play in grounding extrinsic properties (on a
> panprotopsychist view), there may also be the possibility of further
> "direct interactions" between intrinsic properties which would then be

> empirically testable. I guess my question here is what the principled
> difference is between this and standard "extrinsic" interactions.
> After all, when one particle affects another in standard ways, a
> panprotopsychoist can construe this as the intrinsic properties of one
> directly having effects on the intrinsic properties of the other.
>
> Of course physics doesn't "see" the intrinsic part and characterizes
> only the extrinsic structure. But the same would go for Logan's new
> "direct" interactions, it seems to me. Even in parapsychology, one
> could characterize the causal structure extrinsically. All these
> would have an extrinsic structure that an objective science could
> characterize, and all would basic intrinsic properties that are really
> doing the causing and being affected. On either case, the extrinsic
> structure is just a way of getting at the basic causal relations among
> the intrinsic properties. So I'm not sure that I see the difference
> in kind here.

[Logan]

At the time I made the comment that Dave discusses above I had a misunderstanding of what philosophers meant by the term "intrinsic". Now that I have this term clarified in my mind, I believe that I can explain, in a more satisfactory way, what I was trying to get across by the notion of "direct interactions". Dave is absolutely correct in his statement that "the extrinsic structure is just a way of getting at the basic causal relations among the intrinsic properties". Essentially what I am trying to suggest is that proper characterization of the intrinsic properties (with regards to accounting for phenomenal consciousness) may lead to new extrinsic functional relationships among physical primitives that are not predicted by current scientific theory. Such new functional possibilities would presumably arise during the integration of said intrinsic properties with current theory. This possibility can be seen by considering Thompson's discussion of color spaces.

Thompson asks on what basis may we consider Fred to be seeing a new hue as opposed to the possibility that what he is experiencing is not a new hue or color quale, but some other novel form of visual qualia. If I understood his argument correctly, he concluded that we may consider Fred's novel qualitative experience to be that of color if (through relevant psychophysical experimentation) we may determine that the nature of the functional relationships between the new "hue" and the originals are the same as those found in our 3-dimensional color space. Thus Fred's color space is just a 4-dimensional space containing our color space, with the same functional rules of the latter. It is on this basis that we may say that we can have some idea of what Fred's novel experiences are like, since we may use our knowledge of those functional relationships to extrapolate to Fred's experience (although as was discussed in class, we still wouldn't know very much). However if it was determined that Fred's extra qualia did not engage in the same functional relationships with the original elements of our color space, then we can no longer say that what Fred is experiencing is color; it must be some other form of visual qualia (by the way, could it be this type of distinction that Nagel was getting at in his discussion of the qualitative experiences of bats?).

Now in either case the new qualitative feature that Fred is able to perceive may still be considered to be intrinsic in the sense that Dave uses the term in regards to qualia. And in the case that the presence of the new feature merely expands the dimension of the color space without changing the functional/relational rules between the intrinsic elements of the space, we may say that the intrinsic nature of the new "hue" is of essentially the same nature of the primitive elements that make up our color space (i.e. it is a color quale). However in the case that the presence of the new qualitative feature within the space results in a change amongst the functional relationships of the space, then it would seem that the

new feature would have an intrinsic nature substantially different from the elements constituting our color space; it would be such a difference that presumably would be the cause of the change in functionality of the space.

I guess the point that I was trying to make with my last listserve posting was that it may be that the proper characterization of the intrinsic properties (proper in the sense that the characterizations are forced upon us to adequately account for phenomenal consciousness) may lead to functional relationships not predicted by current scientific theory (physical/neuroscientific/psychologic) once such intrinsic properties are integrated into current theory. The existence of such functional relationships may lead to consequences that are empirically testable. Hence the notion of "direct interactions" between intrinsics is really meant to describe possible extrinsic interactions outside the scope of interactions currently postulated by conventional theory.

Logan

From chalmers@paradox.soc-sci.arizona.edu Mon Sep 13 11:56:13 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id LAA06906
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 13 Sep 1999 11:56:13 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFX5U4R08GB8TYJ4@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 13 Sep 1999 11:56:17 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFX5U0VILSB8TH23@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 13 Sep 1999 11:56:11 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id LAA06899 for scicon; Mon, 13 Sep 1999 11:55:58 -0700
Date: Mon, 13 Sep 1999 11:55:58 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Experience as a Kind (from Sarah)
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909131855.LAA06899@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

>From sawright@U.Arizona.EDU Sun Sep 12 21:16:37 1999
Date: Sun, 12 Sep 1999 21:16:39 -0700 (MST)
From: Sarah A Wright <sawright@U.Arizona.EDU>
Subject: Experience as a Kind
To: chalmers@Arizona.EDU

I just wanted to make a comment about an alternate form of Type -A Materialism that I find to be a bit more plausible than the general form the you gloss in Moving Forward on the Problem of Consciousness. In the analogies from Dennett and the Churchlands, the alternate property to be explained seems to be (at least plausibly thought of as) a natural kind. Life, light, and heat, if they are properties to be explained, all seem to be natural kinds. However in one of Dennett's other examples, that of cuteness, the properties admissibility as capturing a natural kind might be brought into question.

Although you have objected to the general schema of using analogies with other properties to undermine the project of identifying consciousness, if you will bear with me, I think that one further analogy may be used to highlight a different objection. This is an analogy with the property (or

at least those things that fall under the description) "In this room". As with those things that fall under the description "What it is like", we have a convenient way to pick out these entities. But having a catch phrase to pick out objects shouldn't lead us to automatically take the grouping as representing a natural kind.

So, let us say that the objects falling under the description "In this room" do not in fact make up a natural kind. We will then absolve our best theories of the world from explaining what it is that gives an object its "In this room"ness; in particular we will not demand that our theories give us a unified explanation of "In this room"ness. Even so, we will want our theories to explain many facts about the individual objects in this room: How my computer works, why my desk weighs as much as it does, etc... One of these individual facts may be: how my desk got into the room. To explain this we do not need to explain the property of "In the room"ness, but rather need to give the history of the desk to trace its path into my house; we want to show how the property of being a desk, and the non-(full fledged)-property of "In the room" came to accidentally occur in the same object.

Ok, so on to "What it is like". Let us say, for the moment that things falling under this description also do not make up a natural kind. If so then, in the same way, our theories do not need to give a unified explanation of "What it is like"ness. Still, of course, we will want to explain many facts about our experiences (those events that fall under the "What it is like"ness) and these are what I take cognitive science to be after. In addition we may want to explain how a particular neural state came to be associated with an experience; in this case we will give the history of the neural state, and how it came, accidentally, to fall under the description "What it is like".**

[[**Also, as an interesting sidenote, if we are actually after a history for each type of experience, we run into further problems. For, as I think Richard Lewontin has argued quite forcefully, there are limitations on the historical explanations available for mental processes. We simply lack the record to make definite judgements about the (long term) history of our mental abilities.]]

An argument based upon this sort of analogy seems to me to fit your Type-A Materialism, but without being eliminativist. Rather than eliminating experiences as things to be explained, this approach simply sets limits on the kinds of explanations we should expect for our experiences.

I don't know if anyone advocates exactly this view, but it seems like a plausible middle road to me.

Sarah

From acboch@U.Arizona.EDU Tue Sep 14 18:12:20 1999
Return-Path: <acboch@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id SAA09371
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 18:12:20 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFYX9V3Y00B8U90M@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 14 Sep 1999 18:12:27 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFYX9U4EPSB8U61U@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Tue, 14 Sep 1999 18:12:26 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id SAA09366 for
<scicon@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 18:12:14 -0700
Received: from localhost (acboch@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id SAA13983 for
<scicon@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 18:12:22 -0700 (MST)
Date: Tue, 14 Sep 1999 18:12:22 -0700 (MST)
From: Adam C Boch <acboch@U.Arizona.EDU>
Subject: Forwarded mail....
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909141811540.12835-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

----- Forwarded message -----
Date: Tue, 14 Sep 1999 13:59:59 -0700 (MST)
From: Adam C Boch <acboch@U.Arizona.EDU>
To: scicon@paradox.soc-sci.arizona.edu

I want to point to an epistemological problem that underlies Chalmers' schema for a science of consciousness now (rather than the third-to-last week when we are scheduled to discuss epistemological issues) since the worry, if serious, seems to undermine the very foundation of the project.

Blind Mary

In class, in the context of altered Mary scenarios, Allen mentioned "blind Mary". It was in passing, and he didn't follow through with the concept, but I want to take the idea and use it to bring out the problem I'm concerned with. Now, it seems to me that a blind Mary scenario (blind, mind you, not just color-blind) would be problematic for the following reason: brain science is reliant on perception (i.e., the perception of the brain scientist(s)). Now it might be argued that any given brain scientist could learn all there is to know regarding the brain even if she were blind (still, I'd imagine there would be problems when she tried to conceptualize the spatio-temporal layout of even the most simple neural networks), but this is not my real worry. My real worry concerns not individual scientists but the nature of the brain sciences themselves. And it seems relatively obvious to me that such sciences rely heavily on perception (observation) for their data. Consider, say, neurophysiology--PET scans, single-cell investigation, brain imaging, exploratory surgery, etc.--without perception there is no data.

The Problem with Perception or What It's Like to be a Human

The problem put short is this: what it's like to be a human might not be what it's like...full stop. In other words, the structure of our experience might not match up as nicely as we might like with the structure of the world. Take Nagel's bat. There is this rich phenomenology--a phenomenological "world" if you will-- a this bat has access to. Now, presumably, it is different (perhaps VERY different) from our own phenomenology (the paper hinges on this point). What it's like for a bat to experience the world is very different from what it's like for a human to experience the world. The world for a human is a world of objects. The human body looks a certain way, feels a certain way, When we observe it, and do so rigorously, adhering to a scientific method, we get things like anatomy and physiology. When we observe the brain at a neuronal level, we get neurophysiology. Now, given that we think a bat's rich phenomenology is different from our own, a bat's neurophysiology (if

it had one) would, presumably, look different. That is, depending how different it is to be a bat rather than a human, that's how different it's scientific picture would be from our own (if you don't like the bat analogy substitute your favorite species, earth-bound or no). But this sounds ridiculous. First, other species do not have science (at least no known one's, we think). Second, we seem to think that, as humans, we are the state of the art of species-making. We are as good as it has gotten so far. That is, we seem to think we stand in some epistemically privileged position with respect to other species. Thus, our science, even if other species' had a science, would be the best. What the world seems like to humans is closer to what the world is actually like than for any other species. The last claim needs some arguing, but it might turn out to be true. Yet even if it is true, it does not get us what we want (or think we already have). That is, we might be the closest of anyone trying to the real story, yet, what it's like to be a human may still not be what it is like.

Kant and the Critique of Chalmers' Proposal for a Science of Consciousness

The noumenal and the phenomenal. Kant thought that no matter how hard we tried, we were stuck in the phenomenal world. All we could ever get was the world as it appeared to humans. The world as it really was, was closed to us. Now, using Kant's distinction as a springboard, I want to make a criticism of Chalmers' proposal for a science of consciousness. Chalmers says that a science of consciousness will consist in something like the following: micro physical facts (facts got by the brain sciences, spec., neurophysiology), linked by bridging laws (perhaps psychophysical) to macro facts (facts about one's own (or human) phenomenology). That is, a science of consciousness will consist in facts about brain processes, or states, linked by laws to facts about human experience. The third person (data from brain observation) adequately linked to the first person (data from experience). The problem is that when it comes time to do the science, i.e., to glean the microphysical facts, we observe the brain. We form theories based on observation. But when we observe the brain we observe it as humans. It is part of what it is like to be a human to have that certain brain-appearance. So the microphysical facts are indexicals--they are facts-for-humans, as it were. In Kantian terms, we get phenomenal facts. So, our science of consciousness ends up linking certain phenomenal facts (i.e., 3rd person ones) with certain OTHER phenomenal facts (i.e., 1st person ones). This isn't what we want. What we want are NEUMENAL facts linked with phenomenal facts. Herein lies explanation. "But I thought Kant said we could never get noumenal facts." is what your thinking. True, but other philosophers after him, those taking the noumena/phenomena distinction seriously, didn't necessarily agree. I don't think we're stuck, and that therefore there can be no science of consciousness as Chalmers' construes it... so more needs to be said about why I think this. However, not today. See my next post (it will still be a pertinent worry as we are now discussing whether or not we can find the neural correlates of consciousness).

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 8 13:20:52 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id NAA06773
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 8 Sep 1999 13:20:52 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFQ9CC70Y0B8TCN4@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 8 Sep 1999 13:20:55 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFQ9CA02M8B8T9S4@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 08 Sep 1999 13:20:53 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id NAA06767 for scicon; Wed, 08 Sep 1999 13:20:47 -0700
Date: Wed, 08 Sep 1999 13:20:47 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Qualia and Knowledge (from Matt)
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909082020.NAA06767@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

>From landsurveyor@hotmail.com Wed Sep 8 10:27:34 1999
Date: Wed, 08 Sep 1999 10:25:43 -0700 (PDT)
From: Matt Herbert <landsurveyor@hotmail.com>
Subject: Qualia and Knowledge
To: chalmers@u.arizona.edu
X-Originating-IP: [128.196.23.108]

Dave,
Sorry to bother you again, but I can't get this sent to the listserv. I
copied the address from your homepage; not sure what the problem is. Feel
free to post to listserv if possible. Thanks
Matt

Jackson's distinction between the physical facts about sensing and the
sensory experience itself is similar in some important ways to Russell's
distinction between knowledge by description and knowledge by acquaintance.
(In turn, Russell and other philosophers probably owe a debt to Aristotle
for suggesting this distinction in Posterior Analytics, but I'm happy to
let the scholars handle that issue.) I'm not sure if Russell's analysis
will comfort the qualia freaks or the reductionists, but it will be worth a
look.

First some background is in order (Cf The Problems of Philosophy, chaps.
4,5). Russell is out to give an exhaustive taxonomy of knowledge, and his
first cut is between knowledge of facts and knowledge of things. This
distinction expresses the difference between knowledge that p and
familiarity with x. The acquaintance/description distinction is a
refinement of the familiarity-with-x concept. I won't attempt to improve on
Russell's clarity regarding this distinction:

"In the preceding chapter we saw that there are two sorts of knowledge:
knowledge of things, and knowledge of truths. In this chapter we shall be
concerned exclusively with knowledge of things, of which in turn we shall
have to distinguish two kinds. Knowledge of things, when it is of the kind
we call knowledge by acquaintance, is essentially simpler than any
knowledge of truths, and is logically independent of knowledge of truths,
though it would be rash to assume that human beings ever, in fact, have
acquaintance with things without at the same time knowing some truth about
them. Knowledge of things by description, on the contrary, always

involves . . . some knowledge of truths as its source and ground" (p. 46).

It may be helpful to think of Mary as lacking acquaintance with a certain quale (say, redness), despite having maximal descriptive knowledge of it. Russell's characterization of acquaintance as logically independent of description matches up with this intuition: one may possess all the descriptions of x and yet not be able to infer what the experience of x is like.

It seems that both Russell and Jackson are making an implicit appeal to the operation of distinct faculties. There are inference modules, whose input can only be facts and whose output can only provide further, extrinsic facts about x's (how they relate to y's, what they do under condition c, etc.), and there are sensory modules, whose input is experience of x and whose output, strictly speaking, is non-propositional, intrinsic information about x's which nonetheless can give rise to propositions to be fed into the inference module.

A faculty of description and a faculty of acquaintance, if you will.

I apologize for not drawing a moral from these bits; it's only a clarifying illustration. I might as well admit my allegiance, however. I think there are good arguments and intuitions in favor of the Russell/Jackson view about qualia. The arguments (that I know) are epistemological and have to do with the need to anchor propositional knowledge in non-propositional experience; the intuitions are those of first-person phenomenology that Dave alludes to in The Conscious Mind.

Matt Herbert
8 Sep 99

Get Your Private, Free Email at <http://www.hotmail.com>

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 8 13:22:34 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id NAA06785
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 8 Sep 1999 13:22:33 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFQ9EFPXXCB8T0WF@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 8 Sep 1999 13:22:36 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFQ9EEQKK0B8TJGY@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 08 Sep 1999 13:22:35 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id NAA06779 for scicon; Wed, 08 Sep 1999 13:22:28 -0700
Date: Wed, 08 Sep 1999 13:22:28 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Epiphenomenalism (from Juraj)
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909082022.NAA06779@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

>From jurosan@usa.net Wed Sep 8 11:35:48 1999
Date: Wed, 08 Sep 1999 20:36:11 +0100 (MET)
From: JURAJ HVORECKY <jurosan@usa.net>
Subject: Epiphenomenalism
To: David Chalmers <chalmers@Arizona.EDU>
X-MIME-Autoconverted: from quoted-printable to 8bit by paradox.soc-sci.arizona.edu id LAA06549

Ok, we have been talking a lot about Jackson yesterday, but we didn't really touch the issue of epiphenomenalism. And there are couple things I want to know about it.

First, I pretty much understand the statement that qualia are epiphenomenal. I suppose my intentional zombie twin riding her car would stop on the red color the very same way I would. And maybe that is what we do - we stop because we process the information about the light and then the thing is somehow ACCOMPANIED by an experience. So there is a slight intuition about it. But epiphenomenalism (E) seems to come in two forms - (E) of mental states and (E) of qualia. Now I do not know much about the first one, but I trust Fodor in his Making minds matter more that is is compatible with physicalism to say the mental events produces causes on the physical.

What I am more interested into is how (E) of qualia is possible. There are two ways to proceed - qualia are inefficacious on the physical and qualia are inefficacious on the mental. It seems like Jackson is switching between the two views. Can qualia be absolutely inefficacious, that is can they have no influence on mental? Well, if that is the case, then they would not be stored in the memory, they would be just passing by. But then the knowledge argument does not work, because they would be NEW at every single moment. In fact, how would one even talk about knowledge, if there is no connection between a quale and the mental event (knowing of x)?

But if they are not epiphenomenal toward other mental states, they might as well have a role in a causal chain toward a physical. If I take for granted that my experiencing of a red light does not cause me to stop (via zombie example), then it still might have a role of comparison this particular red mental image with those previously stored and therefore possibly influence my next zombie-like stop at the red light. To put it in a different way, my qualia might not influence my physical behavior at the moment they are experienced, but leave their guiding trace for later unconscious processing.

So I am quite unsure where exactly (E) comes into the picture.

Juraj

Get free email and a permanent address at <http://www.netaddress.com/?N=1>

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 8 18:24:43 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id SAA07459
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 8 Sep 1999 18:24:43 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFQJY22KQ8B8T6JL@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 8 Sep 1999 18:24:45 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFQJY0QKJKB8T8Y8@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 08 Sep 1999 18:24:44 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id SAA07453 for scicon; Wed, 08 Sep 1999 18:24:38 -0700
Date: Wed, 08 Sep 1999 18:24:38 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Re: Qualia and Knowledge
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909090124.SAA07453@paradox.soc-sci.arizona.edu>
MIME-version: 1.0

Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

On Matt's comments re qualia and knowledge: I think it is right that the difference between knowledge of qualia and knowledge of the external world is closely related to Russell's distinction between knowledge by acquaintance and knowledge by description. Certainly Russell believed that the only true cases of knowledge by acquaintance are cases of direct knowledge of phenomenal states, and he believed that knowledge of the external world was knowledge by description. A lot of contemporary philosophers have disagreed with him about this (e.g. holding that we have perceptual knowledge by acquaintance with external objects), but my own view is that there is something right about his view.

It's certainly true that a "qualia freak" has to pay special attention to the first-person epistemology of qualia, which may well end up being different in interesting ways from the epistemology of the external world. For example, there may be something special to say about the justification of our beliefs about qualia. My own view is that justification of our beliefs in qualia stems at least in part from our acquaintance with the qualia themselves, in a way that doesn't apply to knowledge of external knowledge. Unlike Russell, I'm not sure that I'd call acquaintance a form of knowledge, though; I'd prefer to reserve the latter term for knowledge of truths. Of course like Russell I think that we also have knowledge of truths about qualia that is in some ways grounded in acquaintance.

I talk about this a bit in Chapter 5 of my book, and Tim Bayne has an interesting paper taking issue with some aspects of the view coming out in Philosophy and Phenomenological Research. I've been working on a more developed view in the vicinity of this in a paper called "The Content and Epistemology of Phenomenal Belief", which is online in outline form at

<http://ling.ucsc.edu/~chalmers/papers/belief.html>

I'll probably also give a colloquium in the philosophy dept on this topic in the spring.

I don't think Jackson needs to assume these somewhat controversial views about acquaintance in order for his argument to go through. He needn't make any such theoretical claims about acquaintance, modules, etc, in order to make it plausible that Mary gains new knowledge. Nevertheless, one might well argue that the special character of knowledge of qualia that his thought-experiment exploits stems from the special underlying epistemology.

--Dave.

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 8 18:29:12 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id SAA07468
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 8 Sep 1999 18:29:11 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFQK4LDRB4B8T2C6@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 8 Sep 1999 18:29:15 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFQK4K81XCB8T6LD@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 08 Sep 1999 18:29:13 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu

(8.9.3/8.9.3) id SAA07463 for scicon; Wed, 08 Sep 1999 18:29:07 -0700
Date: Wed, 08 Sep 1999 18:29:07 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Re: Epiphenomenalism
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909090129.SAA07463@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

Juraj raises the issue of epiphenomenalism. I'd prefer not to distinguish qualia from mental states, as qualia are examples of mental states; maybe it's clearer here to distinguish qualia from intentional states such as beliefs. The question Juraj raises is then whether, according to the epiphenomenalist, qualia have causal effects on beliefs. There seem to be at least two possibilities, depending on the status of beliefs:

(1) Qualia cause beliefs, but beliefs don't cause behavior.

(2) Beliefs cause behavior, but qualia don't cause beliefs.

(One might also hold that there is no causal connection between either pair, but I'll set that position aside.)

As far as I recall, Jackson isn't explicit about the causal status of beliefs in his paper, but it's clear from elsewhere that he thinks beliefs are not epiphenomenal and that he is a materialist about beliefs. That seems to commit him to option (2). Some other epiphenomenalists might hold e.g. that beliefs themselves are essentially tied to the phenomenal and so are non-physical, and so would embrace option (1). Both of these are somewhat counterintuitive, especially when it comes to beliefs about qualia. I talk about this issue a fair bit in my book (Chapter 5 again), as I think it's the most serious problem for the epiphenomenalist.

Position (2) seems to raise questions about how we could ever know about qualia, and about how we could ever remember qualia, as Juraj said. In the book I try to defend position (2), or at least to argue that it isn't fatally flawed, partly by arguing that our knowledge of qualia is grounded in our acquaintance with them rather than through any causal connection between qualia and beliefs. In effect the suggestion is that the causal theory of knowledge is inappropriate for knowledge of qualia, for reasons partly to do with those discussed above. One can say something similar about memories. On this view, one would still have the beliefs about the qualia, and the memories; it's just that they wouldn't be directly caused by the qualia.

Of course the view is still somewhat counterintuitive in the way it separates our qualia and our beliefs. I now hold a slightly different view (in the paper mentioned in my reply to Matt; it's also in the book but less emphasized there) on which beliefs about qualia are partly *constituted* by qualia. In effect a quale is a component of a belief about that quale. If that's the case, the relation between qualia and beliefs about qualia is tighter than a mere causal connection; it's a constitutive connection (in effect one is "part" of the other). I think there is strong, independently motivated reasons for believing this in any case. And it gets around the problem about the disconnection between the belief and the quale.

Of course it will turn out that now, on the epiphenomenalist view, beliefs (or at least beliefs about qualia) will be at least partly

nonphysical too, because of the part that is grounded in qualia. Note that one needn't say that they are entirely nonphysical or epiphenomenal. It may be that beliefs are partly constituted by physical/functional states and partly by phenomenal states; if so, their phenomenal part won't have an effect on behavior, but the other part will. I guess something similar *could* apply as a way of making sense of another view Juraj mentions, on which qualia causally affect beliefs which later causally affect behavior. If the aspect of the belief that a quale affects is not the same aspect that causes behavior, then maybe these two things could be reconciled even for an epiphenomenalist.

There are still some awkwardnesses about this position, and all in all I have some preference for the panprotopsychoist view over the epiphenomenalist view partly because of the awkwardnesses, but I do think that epiphenomenalism can at least avoid any fatal flaws here.

--Dave.

From bradt@u.arizona.edu Fri Sep 10 00:01:14 1999
Return-Path: <bradt@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id AAA11625
for <chalmers@paradox.soc-sci.arizona.edu>; Fri, 10 Sep 1999 00:01:14 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFSA0NA0UOB8T2V0@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Fri, 10 Sep 1999 00:01:19 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFSA0LSHUOB8T96J@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Fri, 10 Sep 1999 00:01:17 -0700 (MST)
Received: from snipe.prod.itd.earthlink.net
(snipe.prod.itd.earthlink.net [207.217.120.62]) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) with ESMTP id AAA11620 for <scicon@paradox.soc-sci.arizona.edu>;
Fri, 10 Sep 1999 00:01:07 -0700
Received: from mrector.shiva.com
(ip232.tucson6.az.pub-ip.psi.net [38.29.66.232])
by snipe.prod.itd.earthlink.net (8.9.3/8.9.3) with SMTP id AAA04544 for
<scicon@paradox.soc-sci.arizona.edu>; Fri, 10 Sep 1999 00:01:11 -0700 (PDT)
Date: Fri, 10 Sep 1999 00:05:01 -0700
From: Brad Thompson <bradt@u.arizona.edu>
Subject: RE: Qualia and Knowledge
In-reply-to: <199909082020.NAA06767@paradox.soc-sci.arizona.edu>
To: scicon@paradox.soc-sci.arizona.edu
Reply-to: bradt@u.arizona.edu
Message-id: <000001befb5a\$cd5e72a0\$e8421d26@mrector.shiva.com>
MIME-version: 1.0
X-Mailer: Microsoft Outlook 8.5, Build 4.71.2173.0
Content-type: text/plain; charset=iso-8859-1
Content-transfer-encoding: 7bit
Importance: Normal
X-MSMail-Priority: Normal
X-MimeOLE: Produced By Microsoft MimeOLE V4.72.3110.3
X-Priority: 3 (Normal)
Status: RO

I think that this was an interesting remark that Matt made regarding the knowledge argument and its relationship to Russell's distinction between knowledge by acquaintance and kn. by description:

> It seems that both Russell and Jackson are making an implicit
> appeal to the

- > operation of distinct faculties. There are inference modules,
- > whose input
- > can only be facts and whose output can only provide further,
- > extrinsic facts
- > about x's (how they relate to y's, what they do under condition c, etc.),
- > and there are sensory modules, whose input is experience of x and whose
- > output, strictly speaking, is non-propositional, intrinsic
- > information about
- > x's which nonetheless can give rise to propositions to be fed into the
- > inference module.
- > A faculty of description and a faculty of acquaintance, if you will.

I don't agree that Jackson needs to make such an appeal (though it is certainly not incompatible with his argument and it has an independent plausibility). What I find interesting about the suggestion, though, is that it really points toward one of the stronger materialist defenses against the knowledge argument. Jackson's argument requires that there can be no inference (logically or semantically speaking) from the physical facts to the phenomenal facts. But I think that the Mary thought experiment *by itself* really only demonstrates something weaker--that one cannot make an inference from the physical facts to the phenomenal facts in what one might call a "psychological sense" of inference. And the above psychological story explains why: our representations of sensory states are in a different "format" or a different module (or psychologically encapsulated in some other way) than our so-called descriptive representations of theoretical knowledge.

This is essentially one of Churchland's points in the article we read. But by itself I don't think it fully succeeds against the knowledge argument, since one is still struck by the strong intuition that Mary does in fact learn a new *fact* rather than just acquire a new representation of an old fact. But once again, the above psychological story suggests a materialist solution--one of "explaining away" the apparent distinctness of the two types of facts. It's basically the same point about inferences mentioned above, except now we're not concerned with Mary's psychology but rather with our own psychology (when we are considering the thought experiment). I try to imagine being Mary in the black-and-white room, and so I activate a set of my own descriptive, theoretical representations. This doesn't put me in a state in which my sensory representations of redness are activated (since, according to the present hypothesis, these two types of representations reside in different modules or the like). I then imagine being Mary and seeing a ripe tomatoe--and this *does* activate the relevant sensory representations. This explains why it *seems* to us that Mary learns something new. It would seem that way even if Mary *didn't* learn something new. So the seeming as such can't be taken as evidence for either hypothesis. (This strategy is similar to one given by Christopher Hill in the paper that Dave mentioned before).

We're left now with the question of whether or not we have good reason for thinking that our sensory concept of red is co-referential with some descriptive term in the language of neuroscience (or other science). The above materialist response to the knowledge argument doesn't answer *that* question, but it does suggest that it *could* be for all the knowledge argument shows.

Brad

From chalmers@paradox.soc-sci.arizona.edu Sun Sep 12 22:54:52 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA06004
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 12 Sep 1999 22:54:52 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFWEKCND CWA60B12@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 12 Sep 1999 22:54:55 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFWEKAKYK0B8TD18@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 12 Sep 1999 22:54:52 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id WAA05998 for scicon; Sun, 12 Sep 1999 22:54:42 -0700
Date: Sun, 12 Sep 1999 22:54:42 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: What experience teaches (from Joel M.)
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909130554.WAA05998@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

>From jmartine@U.Arizona.EDU Sun Sep 12 21:33:21 1999
Date: Sun, 12 Sep 1999 21:32:54 -0700 (MST)
From: Joel A Martinez <jmartine@U.Arizona.EDU>
Subject: What experience teaches
To: chalmers@U.Arizona.EDU

In what follows I present some of David Lewis' objections (in "What Experience Teaches") concerning Jackson's Knowledge argument. I also include a brief explanation of the ability hypothesis, as he presents it, along with some of my own comments. (Since I was not sure how many people have read Lewis' paper I included more summary. this made my e-mail longer. I apologize for the length.)

Lewis denies what he calls "The Hypothesis of Phenomenal Information" (HPI) on the grounds that it commits us to a much more "peculiar" view of the world (his words, not mine) than does materialism. But first, let me be clear about what the Hypothesis is; it is familiar to all of us, maybe just not under this name. It is the claim that besides physical information there is an irreducibly different kind of information to be had: *phenomenal information*. So, it is possible for there to be two cases that do not differ physically but do, however, differ, phenomenally. According to Lewis, the characterization of information according to the HPI, is in terms of excluding possibilities. That is, when we get physical information we narrow down the possibilities, sometimes even to one, but there is a range of phenomenal possibilities which we leave open. He uses an analogy of a location of a point on an x-y plane. We may be able to pinpoint the x coordinate exactly, but no amount of x-coordinate info. is going to help us figure out where the y-coordinate lies. "Any amount of x-information leaves open all the y-possibilities." (pg. 583) When we finally do make our y-measurement we acquire a new kind of information, which allows us to eliminate possibilities which were previously open. Similarly, when Mary finally sees red, she is able to eliminate possibilities that were previously left open to her.

Now, invoking the HPI (as a reason why no amount of physical information suffices to teach us what a new experience is like) allows Jackson to construct the knowledge argument. Lewis cannot formulate a version of Materialism (even a very limited version in terms of a small class of possible worlds) which holds given the HPI. He show us that the two are incompatible. As mentioned above, Lewis opts to reject the HPI. He rejects is for two reasons: 1) It is more

"peculiar", 'and therefore less tempting than it may at first seem.' (pg. 586-587) 2) we are not *forced* to accept it. B/c an alternative hypothesis (the ability hypothesis) allows us to explain how experience can best teach us what it is like. That is, we can hold (a) that experience, and nothing else (at least in the actual world) can teach us "what it is like", while still (b) rejecting the HPI and (c) accepting Materialism.

1): Why the HPI is so "peculiar":

First, b/c it is actually opposed to more than just Materialism, according to Lewis. He shows this with the following argument:

"Let parapsychology be the science of all the non-physical things, properties, causal processes, laws of nature, and so forth that may be required to explain the things we do. Let us suppose that we learn ever so much parapsychology. It will make no difference. Black-and -white mary may study all the parapsychology as well as all the psychophysics of color vision, but she still won't know what it is like...our intuitive starting point wasn't just that *physics* lessons could not help the inexperienced to know what it is like. It was that *lessons* couldn't help. If there is such a thing as phenomenal information it isn't just independent of physical information. It's independent of *every sort of information that could be served up in lessons for the inexperienced*. For it is supposed to eliminate possibilities that any amount of lessons leave open. Therefore phenomenal information is not just parapsychological information, if such there be. It's something very much stronger."

(I think there is quite a lot to say about this argument, but I want to comment on the ability hypothesis and, since this is already getting long I will save commentary on this for someone else or for some other time.)

(2) The second reason why HPI is so "peculiar" is that, according to Lewis, the possible alternatives, in the case of something the likes of which we have never experienced (like the alternatives concerning what it would be like to be a bat), is completely unthinkable. Lewis admits that a range of unthinkable alternatives is not impossible; but it is, he says, "peculiar" and is "enough to suggest that we may somehow have gone astray. (I think that the Thompson paper actually suggests something counter to this. It may be possible to get an idea of what some completely foreign experience may be like (e.g. we can plot a novel color by changing the geometry of our color space), and hence generate some range of possibilities. So, the HPI may not be so peculiar, in this respect, after all.)

(3) The third peculiarity is how "isolated" phenomenal information is from all other kinds of info. (I take Lewis here to mean by "isolated", seemingly not making any difference in the physical world.) Mary's behavior would presumably change when she saw green for the first time. to quote Lewis "...her jaw drops and she says "At last! So this is what it's like to see colors!" But, if the HPI is right, she could not be doing any of this as a result of seeing colors for the first time. In order for Mary to do or say anything different her particles must move differently. But how can acquiring phenomenal information make a difference as to the motion of her particles? Lewis makes this point in a somewhat dramatic fashion. "If something non-physical sometimes makes a difference to the motions of physical particles, then physics as we know it is wrong. Not just silent, not just incomplete--wrong." (pg. 590).

The Ability Hypothesis- (Since we discussed this in class, I will say only what is relevant to the comments I make.) So, Lewis sees better reason for rejecting the HPI than materialism. He then (following Nemirow) wants to replace talk of learning a new fact with learning a new ability. Learning what it is like to see a new color is then, like learning to wiggle your ears or eat with chopsticks. No amount of information will suffice for you to acquire the ability. Information might help, but until you do it, you will not know what it is like. Once you know what it is like you have the ability to recall memories of wiggling your ears, imagine wiggling your ears, etc.. So, the ability hypothesis can be formulated as follows "knowing what an experience is like just **is** the possession of these abilities to remember, imagine, and recognize". Lewis' point is that, what enables one to do all this new stuff is not the possession of a new kind of **information**; for Lewis, there are no "phenomenal facts". It is simply **knowing how** to do something.

Commentary:

I think that it would be interesting to look at empirical research that is being done concerning recognition of colors, tastes, etc. in children. I say this b/c when I consider my experience of red or the taste of broccoli, it does not seem to me that I am engaging in the performance of some kind of skill. Red, the experience of it, seems to me to be a given in my subjective experience in the world. But, if it is an acquired ability, then I would have had to develop it, like I developed my ability to use chopsticks. I definitely don't remember developing the ability to see red, but I very well may have when I was very young. I don't know, that is why I think it would be interesting to look at such studies (I am sure there are some out there.). For, if the Ability Hypothesis is right, shouldn't we be able to see this ability develop in children? That is, they would get better at **recognizing** colors; not **naming colors** b/c that could very well concern language acquisition and not sensory experience. Of course, a defender of the ability hypothesis could then say that not all abilities are acquired. Some are natural abilities. This seems plausible to me.

So, the materialist has a response to the question of whether this ability is acquired. But, there remains something unsatisfying about the ability hypothesis. That is, I might be able to see red (and other people cannot) but, explaining **how** I do it still does not seem to me to get around the fact that **I am seeing it**. They still seem to be two different phenomena; the seeing and the ability to see seem distinct to me.

As I have tried to comment above, I also do not think that Lewis has succeeded in making the HPI sound incredibly peculiar. Thompson's paper helps make sense of how we may be able to develop a range of possibilities.

I hope that some find Lewis' arguments interesting. Again, I apologize about the length.

Joel

Joel A. Martinez Social Science Bldg. Rm. 213
Dept. of Philosophy
University of Arizona
P.O. Box 210027

Tucson, AZ 85721-0027
USA

Office #138 Phone 520-621-7098

From anhabib@U.Arizona.EDU Mon Sep 13 00:18:44 1999
Return-Path: <anhabib@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id AAA06132
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 13 Sep 1999 00:18:44 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
id <01JFWHHCYZ3KA60633@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 13 Sep 1999 00:18:48 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDf V5.1-12 #24137) with ESMTP id <01JFWHH9BUQ0B8U7KL@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 13 Sep 1999 00:18:42 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id AAA06127 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 13 Sep 1999 00:18:28 -0700
Received: from localhost (anhabib@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id AAA29879 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 13 Sep 1999 00:18:34 -0700 (MST)
Date: Mon, 13 Sep 1999 00:18:34 -0700 (MST)
From: Allen N Habib <anhabib@U.Arizona.EDU>
Subject: More on the knowledge argument...
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909122245230.9891-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

To all,

I know I'm jumping into the fray a little late, and you might all be heartily sick of hearing about Jackson's knowledge argument, but here are some (few) thoughts about it and the postings concerning it so far.

Just to quickly recap Jackson's initial argument, I think it might be schematised like so:

Premis 1. If materialism is true, then in principle, all mental facts are reducible to physical facts.

Premis 2. If this is the case, then Mary can't learn anything new when she (finally) sees a red tomato, since ex hypothesi Mary already knew all there was to know about the physical facts that might underlay such an experience, i.e. a whole bunch of scientific facts about the workings of the brain.

Premis 3. But Mary certainly seems to learn something new when she sees the tomato. She has a new experience, and all her prior knowledge about the neurology that might underpin such an event is not sufficient to give her this new knowledge without the experience.

Conclusion 1. Not all mental facts are exhausted by physical facts

Conclusion 2. Materialism is false

As Churchland points out, the argument is subject to attack at the second premis, and the charge is one of equivocation. Specifically, two senses of the word 'learn' (or alternatley, the word 'know' if the argument is phrased in those terms) are employed. As Matt

pointed out in his Sept. 8 posting, these two senses might be understood by using Russell's distinction between 'knowledge that' and 'knowledge of' or knowledge of truths and knowledge of things. Matt opines that these terms might correspond to different faculties of the mind: a descriptive faculty and a faculty of acquaintance, respectfully. This would help explain why Mary learns something new, because she receives input to her faculty of acquaintance upon viewing the red tomato, where previously she only had information concerning red in her descriptive faculty.

Dave's reply to Matt (also on Sept 8th) seems to endorse this reading, although Dave is quick to point out that Jackson need not tie himself down to this story of things mental for his argument to go through. All he needs, says Dave, is for Premis three to sound plausible, that is for it to seem to us that Mary does learn something new upon seeing the tomato, regardless of the possible mental ontology underlying it.

But, as Brad points out in his Sept 10th posting, this distinction in types of knowledge can be used against Jackson, especially if something like Matt's 'faculties of the mind' story is adopted. As he (Brad) says ". . .our respresentaions of sensory states are in a different 'format' or a different module (or psychologically encapsulated in some other way) than our so-called descriptive representations of theoretical knowledge.". Brad goes on to say that in fact Mary doesn't learn anything new, but that she is simply put into a state where her sensory representaions of red are activated when she sees the tomato. And, since until she sees the tomato, all she has are the representaions of red to be found in her descriptive faculty, and since those two faculties are insulated from one another, it seems as if she does learn something new. But this intuition (that she does learn something new) is explained by the fact that we can put ourselves into representational states by an effort of the will, whereas it seems that Mary cannot, consequently it seems to us that these latter representations are new or extra information, when in fact they are but different representations of the same information, stored in different formats in the different modules of our mind. (this last bit is just my feeble attempt at a recounting of Brad's argument, and may well be completely incorrect. I would welcome comments and corrections if this is the case, from Brad or anyone)

Here I part ways with Brad, and I think that a materialist might better reply to Jackson that it is because of the equivocation on knowledge or learning in his argument that it seems that Mary 'knows' or 'learns' something new. My argument would go like this: Mary has a neural event(s) that she had never had before upon seeing the red tomato, and insofar as that neural event had causal repercussions on the rest of her brain (i.e. leaving traces in memory, or altering or adding to those parts of the brain that are responsible for what we have here been calling the descriptive faculty) then, if this is what we mean by learning something, Mary has definitely learned something. But this is as it should be, since Mary's brain underwent changes that it had never undergone before. This is at least a necessary condition for learning something new, on a materialists analysis of that term, so it is not surprising that Mary has learned something new when it occurs. The problem is that, in premise two, we are told that Mary already 'knows' everything there is to know about the physical facts that underlay colour-vision, and this is disingenuous. By 'knows' here do we mean that Mary has had the neural events that correlate with every pertinent bit of information about colour-vision? or do we just mean that Mary has studied and memorised (and believes) all the propositions that correspond to facts about the neural correlate(s) of colour-vision. If the latter, then a different sense of knowledge is being employed than the one in the above explanation of how Mary comes to learn something new, since it artificially limits the range of neural events

that are to count as learning to those involved with reading and understanding sentences.

If the former sense is in use, then what reasons do we have for excluding those neural events that underwrite normal colour-vision as pertinent? How can we say truly of Mary that she has all the knowledge about colour-vision there is to have if her brain has never undergone these paradigm transformations? I don't think we can. It is only if we invoke the restricted sense of knowledge, that of sentence appreciation, that this exclusion makes sense, but then of course Jackson's argument has no sting, since all it seems to be saying is that Mary didn't have a certain type of neural event before she saw the tomato, and she did after she saw the tomato, and that the differences in her brain between the two stages qualifies her as having learned something after having seen it.

In other words (if anybody is still with me) I don't think materialists need to deny that Mary learns anything new, they can simply say that in the broad, materialistic sense of knowledge, Mary does learn something new, since her brain acts in ways it never did before, but in this sense of learning, new physical facts are appealed to, namely the events in the brain that underwrite colour-vision, and if we deny 'knowledge' of these, meaning of course that we deny her the actual neural events, and not just sentences describing the neural events, then we can't say that she knows everything that there is to know about colour-vision.

Phееееew. Sorry that took so long, and I also apologise for the inelegant writing, but the material forces some very laboured constructions. Also I'm a terrible writer. Please help me on this or any other score with your comments.

Thanx to all
Allen

"Time flies like an arrow. Fruit
flies like a banana."
Groucho Marx

From rachael@U.Arizona.EDU Mon Sep 13 13:42:04 1999
Return-Path: <rachael@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id NAA07149
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 13 Sep 1999 13:42:03 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFX9JD3GXS8U2IV@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 13 Sep 1999 13:42:08 MST
Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTTP id <01JFX9JASRS0B8U29V@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Mon, 13 Sep 1999 13:42:05 -0700 (MST)
Received: from localhost (rachael@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTTP id NAA05424; Mon,
13 Sep 1999 13:42:03 -0700 (MST)
Date: Mon, 13 Sep 1999 13:42:02 -0700 (MST)
From: Rachael J Parkinson <rachael@U.Arizona.EDU>
Subject: Re: Experience as a Kind (from Sarah)
In-reply-to: <199909131855.LAA06899@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909131244300.999-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII

Status: R0

I have some concerns regarding reduction, particularly as it is laid out by Churchland in "Reduction, Qualia, and the Direct Introspection of Brain States." In this article, Churchland seems to be arguing for a weaker type of reduction, one in which not even indirect deducibility is a requirement. On the D-N model of reduction we see that thermodynamics is reducible: molecule dynamics plus dynamic/thermodynamic bridge laws entail thermodynamics. Churchland weakens this view by arguing that there does not necessarily have to be a smooth reduction. My first question is- how is this view of reduction so different from an epiphenomenalist view that brain facts plus psychophysical bridge laws entail phenomenal facts?

In class, Dave argued that the epiphenomenalist view is not really a *reduction* of experience to brain states. This is because psychophysical laws are not necessary but contingent. Dave further argued that for a strong sense of reduction you need the micro facts to entail the macro facts without an irreducible appeal to laws. This is a view in which P (micro facts) necessarily implies Q (macro facts). Or in other words, in all possible worlds which P holds, Q holds.

But what are the micro facts? They are a combination of microphysical entities and microphysical laws. If you take away the laws, the reduction can not hold. (We can imagine a world where the microphysical entities are the same but the microphysical laws have changed. This world would be very different from our own.) My second question is this: if we throw psychophysical laws into the pot, then couldn't we say that the micro facts (which include various bridge laws, particularly psychophysical bridge laws) necessarily entail the macro facts (including consciousness). So it would be impossible to conceive of a world that had the same micro facts/laws as ours that did not also have the same macro facts, (again, including consciousness.)

I guess my concern can be traced back to Dave's contention that psychophysical laws are not necessary but contingent. It seems to me that most physical laws are not necessary across all possible worlds. The view that brain facts plus psychophysical bridge laws entail phenomenal facts does not seem that different from other scientific reductions, especially given Churchland's weaker requirements for reduction. If someone could please elucidate on the difference between scientific reduction in the realm of physics, for example, and Dave's non-reductive view of consciousness I would appreciate it.

Cheers,
Rachael

From chalmers@paradox.soc-sci.arizona.edu Tue Sep 14 00:16:33 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id AAA08249
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 00:16:33 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFXVP1PB0GB8UIJA@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 14 Sep 1999 00:16:38 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFXVOZ1VU0B8U53R@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Tue, 14 Sep 1999 00:16:35 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id AAA08243 for scicon; Tue, 14 Sep 1999 00:16:24 -0700
Date: Tue, 14 Sep 1999 00:16:24 -0700
From: David Chalmers <chalmers@arizona.edu>

Subject: Comments

To: scicon@paradox.soc-sci.arizona.edu

Message-id: <199909140716.AAA08243@paradox.soc-sci.arizona.edu>

MIME-version: 1.0

Content-type: TEXT/PLAIN; CHARSET=US-ASCII

Status: R

Hi, here are some very brief comments on people's postings:

(1) Re Sarah's comments: The idea that consciousness might not be a natural kind is interesting. Personally I think it is, if anything is; but I'm not sure whether any of these issues really turn on the matter. Even when a property is a very unnatural kind (e.g. "in my desk"), maybe we won't be able to give a unified explanation of its instances, but that won't stop one from giving a fine explanation of any single instance of the property. Here we're not concerned so much with historical explanation (which can be tricky) as reductive explanation -- just giving an account of the micro facts which will necessitate and explain the fact that e.g. a particular pen is in my desk. And that doesn't seem too hard in principle. But for consciousness, even explaining a single instance seems tricky. So I'm not sure the issues are entirely analogous.

(2) Re Allen's comments: I think what really matters is not sentence-knowledge but what we might call factual knowledge, i.e. knowledge of the way the world is. Even if knowing what red is like isn't sentence knowledge, it seems to be factual knowledge. I take it that Allen's central response is to deny that Mary has complete physical knowledge before seeing red, because she hasn't experienced red yet. If that is to apply to factual knowledge (which is what's relevant here), then I suppose it implies that knowing all the truths of physics doesn't give one complete factual knowledge. In a certain sense that might be enough for Jackson already -- if there are facts that go beyond physical facts, he might say he's home.

Perhaps he might concede the point that to know those facts, one has to go into a certain neural state. That would imply there is a close connection between the neural state and the new fact; maybe one could even construe it as a fact about the state (e.g. what it's like to be in that state). But there's nothing here that suggests or implies that it's a *physical* fact about that state, and Jackson would presumably argue that the situation dictates the opposite. So I think the materialist may need to appeal to something more than the mere fact that being in the neural state is required for knowledge here.

(3) Re Rachael's comments: On the entailment model of reductive explanation, the claim will be that the micro facts entail the macro facts. The micro facts here may include the micro laws, but they may not include any special bridge laws. E.g. for chemistry to be reductively explained in terms of physics, on this view, is for the chemical facts to be entailed by the physical facts (including physical laws). But in the case of consciousness, the anti-reductionist's claim is that the physical facts and physical laws don't entail the existence or character of consciousness. If we throw the bridge laws into the base, then we might get an entailment, but that's just what we're not allowed to do where reductive explanation is concerned.

The relevant difference here is that ordinary scientific reductions will have an entailment from micro facts and laws without bridge laws, whereas on the nonreductive view of consciousness, one needs to

include bridge laws to get an entailment. The very different status of micro laws and bridge laws is what makes the difference here.

(4) Re Simon's and Joel M.'s comments: Some good points, to which I don't have much to add here. I think Crick and Koch managed to get a bit confused in the logic of their reasoning about essential features of consciousness, in just the way that Simon nicely points out. Re the psychophysical principles that I give, these are certainly very far from complete, and they certainly don't give sufficient conditions for consciousness. A final theory of consciousness will need principles that are much more detailed and specific.

--Dave.

From press@U.Arizona.EDU Tue Sep 14 01:09:42 1999
Return-Path: <press@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id BAA08363
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 01:09:41 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFXXJY25XCB8TZ2D@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 14 Sep 1999 01:09:47 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFXXJWTV4B8U1Y7@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Tue, 14 Sep 1999 01:09:46 -0700 (MST)
Received: from f1n3.u.arizona.edu
(IDENT:press@f1n3.U.Arizona.EDU [128.196.137.103])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id BAA08358 for
<scicon@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 01:09:36 -0700
Received: from localhost (press@localhost) by f1n3.u.arizona.edu (8.8.8/8.8.8)
with ESMTP id BAA41554 for <scicon@paradox.soc-sci.arizona.edu>; Tue,
14 Sep 1999 01:09:42 -0700
Date: Tue, 14 Sep 1999 01:09:42 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>
Subject: Allen's solution to Mary Arg
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909140100210.49412-1000000@f1n3.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

All-

Allens solution to Jacksons knowledge argument is similar to one that Ive been mulling over for several days now. Actually, it is probably the very same argument. I dont think I can greatly improve on Allens statement of it, so Ill just re-post it before talking about an anticipated objection.

>My argument would go like this: Mary has a neural
>event(s) that she had never had before upon seeing the red tomato, and
>insofar as that neural event had causal reprecussions on the rest of her
>brain (i.e. leaving traces in memory, or altering or adding to those
>parts
>of the brain that are responsible for what we have here been calling the
>descriptive faculty) then, if this is what we mean by learning something,
>Mary has definitely learned something. But this is as it should be, since
>Mary's brain underwent changes that it had never undergone before. This
>is
>at least a necessary condition for learning something new, on a
>materialists analysis of that term, so it is not surprising that Mary has
>learned something new when it occurs. The problem is that, in premise

two,
>we are told that Mary already 'knows' everything there is to know about
>the physical facts that underlay colour-vision, and this is disingenuous.
>By 'knows' here do we mean that Mary has had the neural events that
>correlate with every pertinent bit of information about colour-vision? or
>do we just mean that Mary has studied and memorised (and believes) all
>the
>propositions that correspond to facts about the neural correlate(s) of
>colour-vision. If the latter, then a different sense of knowledge is
being
>employed than the one in the above explanation of how Mary comes to learn
>something new, since it artificially limits the range of neural events
>that are to count as learning to those involved with reading and
>understanding sentences.
> If the former sense is in use, then what reasons do we have for
>excluding those neural events that underwrite normal colour-vision as
>pertinent? How can we say truly of Mary that she has all the knowledge
>about colour-vision there is to have if her brain has never undergone
>these paradigm transformations? I don't think we can. It is only if we
>invoke the restricted sense of knowledge, that of sentence appreciation,
>that this exclusion makes sense, but then of course Jackson's argument
>has
>no sting, since all it seems to be saying is that Mary didn't have a
>certain type of neural event before she saw the tomato, and she did after
>she saw the tomato, and that the differences in her brain between the two
>stages qualifies her as having learned something after having seen it.
> In other words (if anybody is still with me) I don't think
>materialists need to deny that Mary learns anything new,

Just to ensure that we are indeed thinking the same thing, the idea is that learning (coming to know) on the materialist account (and on some dualist accounts too really) just comes down to having a brain state that somehow represents the fact known. If we suppose that materialism is true, Marys experience when she first sees a ripe tomato also comes down to having a particular sort of brain state. So the materialist can claim that though Mary does indeed learn something when she has her first experience of red, he can also claim that she learns a materialistic fact, i.e. what happens when that brain state occurs in her brain.

The objection to this argument that I see looming, and that (I think) Dave alluded to when we were discussing Churchland, is that Allen and I have missed the point - namely that Mary can gain this knowledge ONLY through experiencing it. She can know everything else about color vision through other means, so the experience fact (what it is like) cannot be reduced to these other facts. It seems, the objector might say, that if the experience fact were just one more physical fact, Mary ought to be able to deduce it from the rest of the physical facts she knows. After all, Mary even knows everything there is to know about the very particular brain state that produces red sensations, but she cannot know what it is like to have that brain state in her brain without experiencing it. I suppose Allen and I can still claim that the experience fact is still a physical fact of some sort, but it is apparently a weird sort of non-reductive physical fact, which is about as bad as a non-reductive non-physical fact from a reductivists point of view. At least with non-physical facts there would be an explanation of why the reduction fails.

I think the way around this objection is to show that, while there is a difference between the way Mary knows "what it is like to experience red," and the way that she knows about other physical facts (like which brain state causes subjects to experience red), this difference is not a fundamental difference between two different ways of knowing (a la Matts

reference to Russell). The claim is that all learning just comes down to finding ways of appropriately modifying ones brain, and that all functionally equivalent means of achieving this end count as learning the same thing.

Consider the various ways that Mary can learn about "what it is like to experience red."

- 1) Mary looks at a ripe tomato (or other red object). Light bounces off the tomato, interacts with her retinae, causing neurons to fire, and a particular brain state arises. So long as Mary remembers the salient features of this experience, she knows what it is like to experience redness.
- 2) While still in her black and white room, Mary uses her extensive knowledge of brain physiology and the other sciences to directly stimulate her brain in such a way that it enters the state that she knows should cause her to have an experience of redness.
- 3) While still in her black and white room, Mary uses her extensive knowledge of brain physiology and the other sciences to develop monitoring devices with which she can observe the various processes going on in her brain (like an EEG only much much better). Using the real-time information provided by these devices as biofeedback, she trains herself to manipulate her brain into any state she desires. (I think I read somewhere that some people can manipulate a computer mouse pointer by learning to manipulate certain brain waves. The ability Im attributing to Mary is like that, but much more complex. Mary, of course, has a complete knowledge of her own brain, so she can achieve a higher level of control.) Once she has mastered the control of her brain, (and no longer needs the crutch of the monitoring devices) she decides she wants to experience redness, so she manipulates herself into the appropriate state and learns about this experience.

The point is not that Mary experiences red in 1) but learns about red in 3), but rather that she both learns and experiences in all the cases. The implementation of learning varies greatly, but in each case she comes to know the same fact. 1) is the way most of us learn about redness. In 2) Mary learns about redness without ever seeing a red object. In 3) she learns about red without any external manipulation. All she needs are lots of facts about her brain and practice with controlling it. Once she has really mastered these techniques, when color-deprived Mary thinks, "what is red like?," the answer is just a thought away. But 1), 2), and 3) all produce the same knowledge.

These imagined cases are obviously pretty far-fetched, but then so is Jackson's claim that Mary knows "all the physical facts." But I can make the same point about learning with a more down to earth example.

When we learn basic mathematics, like multiplication say, we use different techniques as we go. When I first learned multiplication we were given worksheets with grids on them. For 4x5 there would be a 4 by 5 grid, and we were supposed to count the 20 squares formed by the grid to get the answer. I thought this was cool, so I used it for quite a while. However, as the numbers got bigger the grids got clumsier and I grudgingly learned to multiply multi-digit numbers on paper, then in my head. Though they differed in implementation, all of these exercises counted as learning the product of two numbers.

Theres probably much more that could be said about this. But not tonight. Let me know what you think.

Joel

PS - I see that Dave has already replied to Allen by the time I've written and posted this. (Do you never sleep?) I only looked at that post briefly, but I think my comments are relevant to Dave's reply. However, since I do sleep, I will let these comments stand as is for now.

From lachter@u.arizona.edu Tue Sep 14 10:09:42 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id KAA08815
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 10:09:42 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFYGFGM1E0B8U7N1@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 14 Sep 1999 10:09:49 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFYGFFHVL8UAM8@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Tue, 14 Sep 1999 10:09:47 -0700 (MST)
Received: from trifid.u.arizona.edu
(IDENT:root@trifid.U.Arizona.EDU [128.196.137.197])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id KAA08810 for
<scicon@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 10:09:36 -0700
Received: from [128.196.99.98] ([128.196.99.98])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id KAA20726; Tue,
14 Sep 1999 10:09:43 -0700
Date: Tue, 14 Sep 1999 10:08:30 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: Allen's solution to Mary Arg
In-reply-to: <Pine.A41.4.10.9909140100210.49412-100000@f1n3.u.arizona.edu>
X-Sender: lachter@pop.u.arizona.edu
To: Joel K Press <press@u.arizona.edu>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a06b4042db2aed5@[128.196.99.98]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
Status: R0

Joel Press asks us to "onsider the various ways that Mary can learn about 'what it is like to experience red.'" But all three of the ways he mentions involve actually experiencing red. It is not clear to me that this is necessary. Suppose Mary knows enough about what her reactions would be to seeing red that she could actually recognize red. This could be fairly automatic given Mary's advanced state of knowledge. Further suppose that she COULD induce the experience of red in herself. Again, this could be fairly automatic since she knows a lot about how the brain works and presumably understands the mental gymnastics involved in biofeedback kind of activities better than anyone currently does. Here is the question, if Mary can recognize red experiences and can induce red experiences in herself in what sense is she missing "facts" about what red is like?

Joel (Lachter)

At 1:09 AM -0700 9/14/99, Joel K Press wrote:

>All-

>

>Allens solution to Jacksons knowledge argument is similar to one that Ive
>been mulling over for several days now. Actually, it is probably the very
>same argument. I dont think I can greatly improve on Allens statement of
>it, so Ill just re-post it before talking about an anticipated objection.

>
>>My argument would go like this: Mary has a neural
>>event(s) that she had never had before upon seeing the red tomato, and
>>insofar as that neural event had causal repercussions on the rest of her
>>brain (i.e. leaving traces in memory, or altering or adding to those
>>parts
>>of the brain that are responsible for what we have here been calling the
>>descriptive faculty) then, if this is what we mean by learning something,
>>Mary has definitely learned something. But this is as it should be, since
>>Mary's brain underwent changes that it had never undergone before. This
>>is
>>at least a necessary condition for learning something new, on a
>>materialists analysis of that term, so it is not surprising that Mary has
>>learned something new when it occurs. The problem is that, in premise
>two,
>>we are told that Mary already 'knows' everything there is to know about
>>the physical facts that underlay colour-vision, and this is disingenuous.
>>By 'knows' here do we mean that Mary has had the neural events that
>>correlate with every pertinent bit of information about colour-vision? or
>>do we just mean that Mary has studied and memorised (and believes) all
>>the
>>propositions that correspond to facts about the neural correlate(s) of
>>colour-vision. If the latter, then a different sense of knowledge is
>being
>>employed than the one in the above explanation of how Mary comes to learn
>>something new, since it artificially limits the range of neural events
>>that are to count as learning to those involved with reading and
>>understanding sentences.
>> If the former sense is in use, then what reasons do we have for
>>excluding those neural events that underwrite normal colour-vision as
>>pertinent? How can we say truly of Mary that she has all the knowledge
>>about colour-vision there is to have if her brain has never undergone
>>these paradigm transformations? I don't think we can. It is only if we
>>invoke the restricted sense of knowledge, that of sentence appreciation,
>>that this exclusion makes sense, but then of course Jackson's argument
>>has
>>no sting, since all it seems to be saying is that Mary didn't have a
>>certain type of neural event before she saw the tomato, and she did after
>>she saw the tomato, and that the differences in her brain between the two
>>stages qualifies her as having learned something after having seen it.
>> In other words (if anybody is still with me) I don't think
>>materialists need to deny that Mary learns anything new,
>
>Just to ensure that we are indeed thinking the same thing, the idea is
>that learning (coming to know) on the materialist account (and on some
>dualist accounts too really) just comes down to having a brain state that
>somehow represents the fact known. If we suppose that materialism is
>>true, Marys experience when she first sees a ripe tomato also comes down
>to having a particular sort of brain state. So the materialist can claim
>that though Mary does indeed learn something when she has her first
>experience of red, he can also claim that she learns a materialistic fact,
>i.e. what happens when that brain state occurs in her brain.
>
>The objection to this argument that I see looming, and that (I think) Dave
>alluded to when we were discussing Churchland, is that Allen and I have
>missed the point - namely that Mary can gain this knowledge ONLY through
>experiencing it. She can know everything else about color vision through
>other means, so the experience fact (what it is like) cannot be reduced to
>these other facts. It seems, the objector might say, that if the
>experience fact were just one more physical fact, Mary ought to be able to
>deduce it from the rest of the physical facts she knows. After all, Mary

>even knows everything there is to know about the very particular brain
>state that produces red sensations, but she cannot know what it is like to
>have that brain state in her brain without experiencing it. I suppose
>Allen and I can still claim that the experience fact is still a physical
>fact of some sort, but it is apparently a weird sort of non-reductive
>physical fact, which is about as bad as a non-reductive non-physical fact
>from a reductivists point of view. At least with non-physical facts there
>would be an explanation of why the reduction fails.

>
>I think the way around this objection is to show that, while there is a
>difference between the way Mary knows "what it is like to experience red,"
>and the way that she knows about other physical facts (like which brain
>state causes subjects to experience red), this difference is not a
>fundamental difference between two different ways of knowing (a la Matts
>reference to Russell). The claim is that all learning just comes down to
>finding ways of appropriately modifying ones brain, and that all
>functionally equivalent means of achieving this end count as learning the
>same thing.

>
>Consider the various ways that Mary can learn about "what it is like to
>experience red."

>
>1) Mary looks at a ripe tomato (or other red object). Light bounces off
>the tomato, interacts with her retinae, causing neurons to fire, and a
>particular brain state arises. So long as Mary remembers the salient
>features of this experience, she knows what it is like to experience
>redness.

>
>2) While still in her black and white room, Mary uses her extensive
>knowledge of brain physiology and the other sciences to directly stimulate
>her brain in such a way that it enters the state that she knows should
>cause her to have an experience of redness.

>
>3) While still in her black and white room, Mary uses her extensive
>knowledge of brain physiology and the other sciences to develop monitoring
>devices with which she can observe the various processes going on in her
>brain (like an EEG only much much better). Using the real-time
>information provided by these devices as biofeedback, she trains herself
>to manipulate her brain into any state she desires. (I think I read
>somewhere that some people can manipulate a computer mouse pointer by
>learning to manipulate certain brain waves. The ability Im attributing to
>Mary is like that, but much more complex. Mary, of course, has a complete
>knowledge of her own brain, so she can achieve a higher level of control.)
>Once she has mastered the control of her brain, (and no longer needs the
>crutch of the monitoring devices) she decides she wants to experience
>redness, so she manipulates herself into the appropriate state and learns
>about this experience.

>
>The point is not that Mary experiences red in 1) but learns about red in
>3), but rather that she both learns and experiences in all the cases. The
>implementation of learning varies greatly, but in each case she comes to
>know the same fact. 1) is the way most of us learn about redness. In 2)
>Mary learns about redness without ever seeing a red object. In 3) she
>learns about red without any external manipulation. All she needs are
>lots of facts about her brain and practice with controlling it. Once she
>has really mastered these techniques, when color-deprived Mary thinks,
>"What is red like?" the answer is just a thought away. But 1), 2), and
>3) all produce the same knowledge.

>
>These imagined cases are obviously pretty far-fetched, but then so is

>Jacksons claim that Mary knows "all the physical facts." But I can make
>the same point about learning with a more down to earth example.
>
>When we learn basic mathematics, like multiplication say, we use different
>techniques as we go. When I first learned multiplication we were given
>worksheets with grids on them. For 4x5 there would be a 4 by 5 grid, and
>we were supposed to count the 20 squares formed by the grid to get the
>answer. I thought this was cool, so I used it for quite a while.
>However, as the numbers got bigger the grids got clumsier and I grudgingly
>learned to multiply multi-digit numbers on paper, then in my head. Though
>they differed in implementation, all of these exercises counted as
>learning the product of two numbers.
>
>Theres probably much more that could be said about this. But not tonight.
>Let me know what you think.
>
>Joel
>
>PS - I see that Dave has already replied to Allen by the time I've written
>and posted this. (Do you never sleep?) I only looked at that post
>briefly, but I think my comments are relevant to Dave's reply. However,
>since I do sleep, I will let these comments stand as is for now.

From press@U.Arizona.EDU Tue Sep 14 21:37:00 1999
Return-Path: <press@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA09819
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 21:36:59 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFZ4FLSJ1SB8Q3ZS@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 14 Sep 1999 21:37:07 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFZ4FKSUDSB8UCHZ@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Tue, 14 Sep 1999 21:37:06 -0700 (MST)
Received: from f1n7.u.arizona.edu
(IDENT:press@f1n7.U.Arizona.EDU [128.196.137.107])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA09814 for
<scicon@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 21:36:51 -0700
Received: from localhost (press@localhost) by f1n7.u.arizona.edu (8.8.8/8.8.8)
with ESMTP id VAA38934; Tue, 14 Sep 1999 21:36:59 -0700
Date: Tue, 14 Sep 1999 21:36:59 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>
Subject: More Mary
To: lachter@U.Arizona.EDU
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909142115170.37568-100000@f1n7.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

All,

Joel Lachter's comments about my comments about Allen's solution to the
Mary Argument are, I think, in the same spirit.

>Joel Press asks us to "onsider the various ways that Mary can learn about
>'what it is like to experience red.'" But all three of the ways he
>mentions
>involve actually experiencing red. It is not clear to me that this is
>necessary. Suppose Mary knows enough about what her reactions would be to

>seeing red that she could actually recognize red. This could be fairly
>automatic given Mary's advanced state of knowledge. Further suppose that
>she COULD induce the experience of red in herself. Again, this could be
>fairly automatic since she knows a lot about how the brain works and
>presumably understands the mental gymnastics involved in biofeedback kind
>of activities better than anyone currently does. Here is the question, if
>Mary can recognize red experiences and can induce red experiences in
>herself in what sense is she missing "facts" about what red is like?

Both of these additional ways of learning about red seem plausible to me if we assume Jackson's stipulation that Mary knows everything about brains, science and so on. We could add these to the list I started as numbers 4) and 5). However, the reason that all my examples contained an experiencing of red is that those who share Jackson's intuitions probably won't accept Joel L.'s cases as real possibilities. My cases were designed to show that experiencing red and other ways of coming to know about red blend into each other, leaving Jackson no reasonable place to draw the line between the two. Joel L.'s cases nicely extend the spectrum of cases.

Thanks for the input.
Joel (Press)

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 15 00:03:22 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTMP id AAA10200
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 15 Sep 1999 00:03:21 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFZ9J2YR34B8U0J3@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 15 Sep 1999 00:03:29 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTMP id <01JFZ9J0UWOWB8UCDP@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 15 Sep 1999 00:03:26 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id AAA10194; Wed, 15 Sep 1999 00:03:13 -0700
Date: Wed, 15 Sep 1999 00:03:13 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Re: Allen's solution to Mary Arg
To: lachter@u.arizona.edu, press@u.arizona.edu
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909150703.AAA10194@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

Quick comments on messages from Joel and Joel (inc.):

(1) Re Joel P: I think Jackson might say the same thing I said to Allen. In a certain sense it might be a "materialistic fact" in that it's a fact about what it's like to be in a given brain state, but on the face of it this still involves different properties of the state in question, properties distinct from the physical properties she knows about already. It may be true that to know the fact in question one has to be in the brain state, but it's not clear that that suffices to make the fact a physical fact. Compare: I may have to be in a certain physical state to know that Paris is in France, but that doesn't mean the Paris fact is a fact about my brain state.

(2) Re Joel L.: Joel suggests that on the basis of her physical knowledge, Mary could learn to indirectly recognize red experiences (by noticing her reactions) and could learn to induce the experiences

in herself. The question then, in what sense is she missing "facts" about what red is like? This is complex, but on the face of it there are lots of cases where a capacity for indirectly recognizing X and constructing X's leaves us mostly ignorant about Xs. Just say we know that fire always cause smoke, and that matches always cause fire, and that's all we know about fire (we've never even seen it). Then we might be in a position where we could recognize the presence of fire (by observing smoke) and where we could bring about the presence of fire (by lighting matches), while knowing hardly anything about the character of fire. We'd just know about one or two associated causal relations. So indirect recognition and construction seem to be fairly weak criteria for knowledge.

--Dave.

From lachter@u.arizona.edu Wed Sep 15 16:41:07 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Aurora.Telcom.arizona.edu [128.196.128.236])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA11677
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 15 Sep 1999 16:41:06 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JG08DN443KA5ZQKS@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 15 Sep 1999 16:41:09 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JG08DHF368B8UJ46@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 15 Sep 1999 16:40:45 -0700 (MST)
Received: from trifid.u.arizona.edu
(IDENT:root@trifid.U.Arizona.EDU [128.196.137.197])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA11672 for
<scicon@paradox.soc-sci.arizona.edu>; Wed, 15 Sep 1999 16:40:27 -0700
Received: from [128.196.99.98] ([128.196.99.98])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id QAA17080; Wed,
15 Sep 1999 16:40:36 -0700
Date: Wed, 15 Sep 1999 16:39:48 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: Allen's solution to Mary Arg
In-reply-to: <199909150703.AAA10194@paradox.soc-sci.arizona.edu>
X-Sender: lachter@pop.u.arizona.edu
To: David Chalmers <chalmers@Arizona.EDU>
Cc: press@u.arizona.edu, scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a00b40567c47fc2@[128.196.99.98]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
Status: RO

Dave,

I don't think I was clear enough about how this example is supposed to work. I think you can automatically and instantaneously recognize something you have never seen before. That was definitely the experience I had when I heard that bat in my bedroom. On a fairly regular basis I recognize things that I have never seen before without doing any "conscious" (nonautomatic) reasoning. I bet when you were first introduced to your girlfriend's parents (assuming such an even has occurred), you recognized her father/mother/brother/dog without having to think "Hmmm, this person is male and about 25 yrs older than her, while this other person it about her age so this must be the father..." It is just automatic when you have enough background information. There are lots of things that people recognize instantly without having seen them. Similarly for imagining things.

Now it seems to me quite plausible that Mary can recognize red automatically. That is, it is not clear to me that it would be necessary for Mary to do any conscious reasoning to see red. Further it is pretty clear that you and I go through unconscious reasoning in recognizing color so any distinction is going to have to lie with the conscious reasoning.

In essence I think that it is quite plausible to think that Mary is just like you and me except that she is not having a red experience and she has no memories of having had such an experience. I don't see how either of those is going to buy you an absence of knowledge.

I also don't like your fire example. It seems to me that in the fire example I know almost nothing about fire, while Mary knows everything (or nearly everything) about red. Similarly, the average person knows much more about fire than that it produces smoke, while the average person knows almost nothing about red except that it causes red sensations. It seems to me that a better analogy might be to compare someone who has only seen flames (say he has seen candles burn), but otherwise knows nothing about fire, to someone who has read extensively about fires but has never actually seen a flame. Who will more quickly identify a forest fire?

Joel

At 12:03 AM -0700 9/15/99, David Chalmers wrote:

>Quick comments on messages from Joel and Joel (inc.):

>(2) Re Joel L.: Joel suggests that on the basis of her physical

>knowledge, Mary could learn to indirectly recognize red experiences

>(by noticing her reactions) and could learn to induce the experiences

>in herself. The question then, in what sense is she missing "facts"

>about what red is like? This is complex, but on the face of it there

>are lots of cases where a capacity for indirectly recognizing X and

>constructing X's leaves us mostly ignorant about Xs. Just say we know

>that fire always cause smoke, and that matches always cause fire, and

>that's all we know about fire (we've never even seen it). Then we

>might be in a position where we could recognize the presence of fire

>(by observing smoke) and where we could bring about the presence of

>fire (by lighting matches), while knowing hardly anything about the

>character of fire. We'd just know about one or two associated causal

>relations. So indirect recognition and construction seem to be fairly

>weak criteria for knowledge.

>

>--Dave.

From lan@U.Arizona.EDU Sat Sep 18 15:03:16 1999

Return-Path: <lan@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id PAA00777

for <chalmers@paradox.soc-sci.arizona.edu>; Sat, 18 Sep 1999 15:03:16 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JG4QJKBSRKB8U0TK@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Sat, 18 Sep 1999 22:04:12 MST

Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

with ESMTP id <01JG4QJIA200B8U1LW@Telcom.Arizona.EDU> for

chalmers@Arizona.EDU; Sat, 18 Sep 1999 22:04:10 -0700 (MST)

Received: from localhost (lan@localhost)

by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id WAA17291; Sat,

18 Sep 1999 22:04:08 -0700 (MST)

Date: Sat, 18 Sep 1999 22:04:08 -0700 (MST)

From: Lonnie A Nelson <lan@U.Arizona.EDU>

Subject: Re: Allen's solution to Mary Arg

In-reply-to: <v04020a00b40567c47fc2@[128.196.99.98]>
To: Joel Lachter <lachter@U.Arizona.EDU>
Cc: David Chalmers <chalmers@Arizona.EDU>, press@U.Arizona.EDU,
scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909182155010.15049-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

Regarding all of the previous... A somewhat annoying trait that I have is that I like to know what I am referring to before I refer to it, therefore, I was wondering, Would someone (anyone) mind defining exactly what is meant by "knowledge", specifically that which mary does or does not have. It seems to me that a person can mean various things when they use the word knowledge.

If it is being used in the "knowledge about the things in the world" sense, I would like a definition (operational, of course) of the word "things" in this sense. It seems to me that an experience is a "thing". Unless you are talking about "things composed of known forms of matter".

So, how is "Knowledge" meant?

--Lonnie

It is a common fate of all knowledge to begin as heresy and end as orthodoxy.

-Thomas Huxley

Lonnie A Nelson
Department of Psychology
Human Energy Systems Laboratory
University of Arizona
lan@u.arizona.edu

From serobert@U.Arizona.EDU Sun Sep 12 22:24:41 1999
Return-Path: <serobert@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA05844
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 12 Sep 1999 22:24:40 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFWDHXESJ4B8TS8F@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 12 Sep 1999 22:24:43 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFWDHU6H40B8TRD1@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 12 Sep 1999 22:24:38 -0700 (MST)
Received: from f1n2.u.arizona.edu
(IDENT:serobert@f1n2.U.Arizona.EDU [128.196.137.102])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA05839 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 12 Sep 1999 22:24:28 -0700
Received: from localhost (serobert@localhost)
by f1n2.u.arizona.edu (8.8.8/8.8.8) with ESMTP id WAA24248 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 12 Sep 1999 22:24:34 -0700
Date: Sun, 12 Sep 1999 22:24:34 -0700 (MST)
From: Simon E Roberts-Thomson <serobert@U.Arizona.EDU>
In-reply-to: <Pine.HPX.4.10.9909111549300.6567-100000@orion.U.Arizona.EDU>
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909122105350.9957-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Dear all,

Given that at this stage we are going on to talk about the neural correlates of consciousness, it may be useful to make clear a relatively simple distinction - that between necessary and sufficient causes. The reason why I think that this is important is that, whilst reading the Crick and Koch article, it occurred to me that they seem to have made the mistake of conflating the two.

Specifically, I would refer you to their 1998 article (in this weeks reading), where they are talking about those aspects of the discussion about consciousness that they have deliberately decided to leave out. In (2), they say:

"It is plausible that some species of animals ... possess some of the essential features of consciousness, but not necessarily all. For this reason, appropriate experiments on such animals may be relevant to finding the mechanisms underlying consciousness. It follows that a language system (of the type found in humans) is not essential for consciousness -- that is, one can have the key features of consciousness without language."

This passage seems rather confused. It seems that they are suggesting that since some animals both have some essential features of consciousness and do not have language, then language is not essential for consciousness. What, however, do they mean by "essential"? If they mean that higher mammals have some of the cognitive mechanisms that are *necessary* for consciousness, then their conclusion does not follow - there is no reason to suppose that language is not necessary for consciousness as we understand it. Whilst those animals might possess some of the necessary features of consciousness, it is at least uncertain whether those features are *sufficient* for them to be considered conscious. If on the other hand they are saying that these animals have cognitive features that are sufficient for consciousness, then why do they say that these animals only possess some, but not all, of the "essential"

features of consciousness? It would seem that by definition that these animals would have all the cognitive faculties that are required for consciousness.

The importance of this distinction may perhaps be made clearer by applying it to a familiar example - Dave's psychophysical principles. If these three principles are merely necessary for connecting processing and experience, then we need to ask the further question of whether they are also sufficient. If the answer is no, then we need to search for further principles that are necessary and which, in combination with the other three, will be sufficient principles.

I hope this has helped

Simon.

From chalmers@paradox.soc-sci.arizona.edu Wed Sep 15 01:14:35 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id BAA10428
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 15 Sep 1999 01:14:35 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JFZC1D9JLCB8UQ52@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 15 Sep 1999 01:14:43 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JFZC1BPNTSB8UMDX@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 15 Sep 1999 01:14:40 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id BAA10419 for scicon; Wed, 15 Sep 1999 01:14:28 -0700
Date: Wed, 15 Sep 1999 01:14:28 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: brain pictures
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909150814.BAA10419@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

Hi all, we were discussing (in class and in break) the frontal lobe, prefrontal cortex, etc. There's a nice simple diagram at

<http://www.waiting.com/frontallobe.html>

It seems that the frontal lobe consists of prefrontal cortex, premotor cortex, and motor cortex (plus the frontal eye fields which are distinct on some accounts). The prefrontal cortex is the front part of the lobe (all the way to the very front), as a couple of us thought. The premotor and motor areas are behind that.

You can get from that page to another simple picture of areas of the brain, if that's helpful.

There is some info (and diagrams) on the differences between non-human primate vs. human brains at

<http://psy.anu.edu.au/unit/b07/OnWeb/B07Lect02-99.htm>

You can get to lots of other pictures of the brain by going to

<http://dir.yahoo.com/Science/Biology/Anatomy/Brain/>

Finally, it turns out the Sheinberg and Logothetis reading for next week is on the web. I've put in a link to it from the main web page for the seminar.

--Dave.

From lachter@u.arizona.edu Tue Sep 14 10:52:05 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id KAA08876
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 14 Sep 1999 10:52:05 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
id <01JFYHW17TN4B8U5E1@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 14 Sep 1999 10:52:12 MST
Received: from trifid.u.arizona.edu by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
with ESMTP id <01JFYHVZYXJ4B8U2P5@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Tue, 14 Sep 1999 10:52:10 -0700 (MST)
Received: from [128.196.99.98] ([128.196.99.98])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id KAA28596; Tue,
14 Sep 1999 10:52:09 -0700
Date: Tue, 14 Sep 1999 10:50:49 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: Comments
In-reply-to: <199909140716.AAA08243@paradox.soc-sci.arizona.edu>
X-Sender: lachter@pop.u.arizona.edu
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a07b40431caa6cb@[128.196.99.98]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
Status: RO

Okay, I haven't been saving people's mail and I can't remember exactly what Sarah wrote, but Dave's reply resonated with something I didn't like in his paper from this week's readings. Thus I will reply even if this is off the thread.

In the readings for this week both Chalmers and Block seem to assume that it is possible to find a neural correlate of consciousness. This struck me as odd. It seems to me that we are about as likely to find a neural correlate of consciousness as we are to find a silicon correlate of word processing. Lets face it, everything in your computer is involved in word processing, and everything that is involved in word processing is also involved in functions which are not word processing. It seems to me that consciousness is likely to be something like this. In particular, I don't see any reason to think that consciousness is a natural kind from the point of view of neuroscience or cognitive psychology.

So, Question: Is it necessary for consciousness to be a natural kind from the point of view of neuroscience (cognitive psychology) for their to be a neural (cognitive) correlate of consciousness?

Whether or not it is necessary, it is pretty clear that people who are in the business of looking for correlates of consciousness are looking for natural kinds from their own scientific point of view. Look at the list which begins Dave's paper for this week. People just are not proposing correlates which involve complex conditions and interactions between brain/cognitive states.

The point here is that whether or not we think of consciousness as a

natural kind (from the point of view of "science") has a big influence on how we carry out the enterprise of creating a science of consciousness.

Joel

At 12:16 AM -0700 9/14/99, David Chalmers wrote:

>Hi, here are some very brief comments on people's postings:

>

>(1) Re Sarah's comments: The idea that consciousness might not be a
>natural kind is interesting. Personally I think it is, if anything
>is; but I'm not sure whether any of these issues really turn on the
>matter. Even when a property is a very unnatural kind (e.g. "in my
>desk"), maybe we won't be able to give a unified explanation of its
>instances, but that won't stop one from giving a fine explanation of
>any single instance of the property. Here we're not concerned so much
>with historical explanation (which can be tricky) as reductive
>explanation -- just giving an account of the micro facts which will
>necessitate and explain the fact that e.g. a particular pen is in my
>desk. And that doesn't seem too hard in principle. But for
>consciousness, even explaining a single instance seems tricky. So I'm
>not sure the issues are entirely analogous.

From acboch@U.Arizona.EDU Mon Sep 27 19:36:45 1999

Return-Path: <acboch@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id TAA19637

for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 19:36:45 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JGH612NM20B8W5KI@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Mon, 27 Sep 1999 19:36:56 MST

Received: from kitts.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

with ESMTP id <01JGH611LYRKB8W7NJ@Telcom.Arizona.EDU> for

chalmers@Arizona.EDU; Mon, 27 Sep 1999 19:36:55 -0700 (MST)

Received: from localhost (acboch@localhost)

by kitts.u.arizona.edu (8.8.8/8.8.8) with ESMTP id TAA61940 for

<chalmers@arizona.edu>; Mon, 27 Sep 1999 19:36:52 -0700

Date: Mon, 27 Sep 1999 19:36:52 -0700 (MST)

From: Adam C Boch <acboch@U.Arizona.EDU>

Subject: Re:NCC's

In-reply-to: <199909050049.RAA05601@paradox.soc-sci.arizona.edu>

To: David Chalmers <chalmers@Arizona.EDU>

Message-id: <Pine.A41.4.10.9909271228270.53204-100000@nevis.u.arizona.edu>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: RO

Dave,

I thought I'd post directly to you since I seem to take a radically different angle than you and the others on the topics we've covered so far. The assigned readings just seem to miss the mark, in my opinion, so I'm using some outside sources to inform my posting.

Here I'm using the article by Depraz in The View From Within, and Varela's Neurophenomenology paper from the penultimate week's reading list. Perhaps, after reading the post, you could point me to more like this (if there is any).

The problem I have with current NCC research projects is a

methodological/epistemological one. The criticism is inspired by the Husserlian notion of reductions. A recent trend in first-person approaches insists that it is only within a Husserlian methodological framework that we will be able to get rigorous enough about first-person conscious experiential data such that a Chalmersian science of consciousness project will be possible. That is, the claim is that it is only by practicing Husserlian reductions (the epoche and eidetic) that researchers will be able to rigorously gather relevant 1st person data such that such data can successfully be bridged with 3rd person neuroscientific data (among other 3rd person data). The problem is this. Such reductions aim, by definition, at uncovering the essential (irreducible) structure of conscious experience (see Varela p.4 of web print-out). That is, the thought is that what we get when we casually look at the world is only one possible "presentation" (or representation), one we have in virtue of a habitual way of looking at the world. If we bracket this habitual mode, allowing room for other possible presentations, we go from contingent "facts" to essential facts (i.e., we go from facts that are contingent in virtue of being derived from a single perspective to those that are universal in virtue of being derived from a multitude of perspectives--see Depraz p.101).

Now, Varela writes:

We are similarly assuming that human experience follows fundamental structural principals which...enforces the nature of what is given to us as contents of experience (p.8)

and

It is one of the most impressive discoveries of the phenomenological movement to have quickly realized that an investigation of the structure of human experience inevitably induces a shift towards considering several levels of my consciousness as inextricably linked to those of others and to the phenomenal world in an emphatic mesh. Consequently, the usual opposition of first-person vs. third-person accounts is misleading. It makes us forget that so-called third-person, objective accounts are done by a community of concrete people who are embodied in their social and natural worlds as much as first-person accounts... (p.9).

In other words, the relevant phenomenological reductions aim at irreducible structures of not only the act of consciousness but of the objects of consciousness as well. Depraz uses the example of the experience of red (p.101). After performing the reductions one is left with the essential structure of an experience of red--presumably, one that is much richer than the one had previous to the reduction. In other words, one gets essential (or the essence of) red (as an object of consciousness), and this is DIFFERENT (and epistemically privileged) from contingent red (red as an object of consciousness before the reductions are performed).

The result seems to be that, after one performs these phenomenological reductions, one is left with a more robust and fuller phenomenal world. The world looks different, feels different. And--here's the meat of it-- if a neuroscientist performs these reductions, the BRAIN will look and feel different, perhaps quite different. The essential structure of an experience of the brain might be altogether different from the one we have from habitual observation. For example, the way neurons look and seem to behave before the reductions might radically differ from the way they look and seem to behave after--if what we get after can even be called a neuron any longer (i.e., things might look VERY different!).

With a more robust phenomenal world we would presumably get a more robust neuroscience (since most of neuroscientific data is predicated on observation). After some progress on the first-person side of the project, it would seem necessary that, with this honed observational tool, we revamp the third-person side. Only then could we be said to be on our way to a RIGOROUS science of consciousness (after all, this was Husserl's goal).

Of course, this is only a criticism of someone interested in a science of consciousness who takes Husserlian phenomenology seriously as a means of gathering first-person data (specifically, someone like Varela). If one doesn't take Husserl seriously, one doesn't have to take the criticism seriously. As someone who does take Husserl seriously, it seems that, given the above, it is from the first-person perspective that a science of consciousness must begin. The task of gathering third-person data follows only after significant work has been done on the first-person side. Work done in the third-person beforehand, is, if Husserl is right, haphazard. It may be science, but not, as Husserl says, rigorous science.

From acboch@U.Arizona.EDU Mon Sep 27 23:36:39 1999
Return-Path: <acboch@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA20151
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 23:36:39 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGHEEISHTSB8W0LY@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 27 Sep 1999 23:36:50 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGHEEHXZ4B8LRBU@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 27 Sep 1999 23:36:49 -0700 (MST)
Received: from nevis.u.arizona.edu (IDENT:acboch@shell [128.196.137.19])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA20146 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 23:36:11 -0700
Received: from localhost (acboch@localhost)
by nevis.u.arizona.edu (8.8.8/8.8.8) with ESMTP id XAA49114 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 23:36:24 -0700
Date: Mon, 27 Sep 1999 23:36:23 -0700 (MST)
From: Adam C Boch <acboch@U.Arizona.EDU>
Subject: Re:NCC's (fwd)
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909272334360.52474-100000@nevis.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

----- Forwarded message -----
Date: Mon, 27 Sep 1999 19:36:52 -0700 (MST)
From: Adam C Boch <acboch@U.Arizona.EDU>
To: David Chalmers <chalmers@arizona.edu>
Subject: Re:NCC's

All,

The assigned readings just seem to miss the mark, in my opinion, so I'm using some outside sources to inform my posting.

Here I'm using the article by Depraz in The View From Within, and Varela's Neurophenomenology paper from the penultimate week's reading list.

The problem I have with current NCC research projects is a methodological/epistemological one. The criticism is inspired by the Husserlian notion of reductions. A recent trend in first-person approaches insists that it is only within a Husserlian methodological framework that we will be able to get rigorous enough about first-person conscious experiential data such that a Chalmersian science of consciousness project will be possible. That is, the claim is that it is only by practicing Husserlian reductions (the epoche and eidetic) that researchers will be able to rigorously gather relevant 1st person data such that such data can successfully be bridged with 3rd person neuroscientific data (among other 3rd person data). The problem is this. Such reductions aim, by definition, at uncovering the essential (irreducible) structure of conscious experience (see Varela p.4 of web print-out). That is, the thought is that what we get when we casually look at the world is only one possible "presentation" (or representation), one we have in virtue of a habitual way of looking at the world. If we bracket this habitual mode, allowing room for other possible presentations, we go from contingent "facts" to essential facts (i.e., we go from facts that are contingent in virtue of being derived from a single perspective to those that are universal in virtue of being derived from a multitude of perspectives--see Depraz p.101).

Now, Varela writes:

We are similarly assuming that human experience follows fundamental structural principals which...enforces the nature of what is given to us as contents of experience (p.8)

and

It is one of the most impressive discoveries of the phenomenological movement to have quickly realized that an investigation of the structure of human experience inevitably induces a shift towards considering several levels of my consciousness as inextricably linked to those of others and to the phenomenal world in an emphatic mesh. Consequently, the usual opposition of first-person vs. third-person accounts is misleading. It makes us forget that so-called third-person, objective accounts are done by a community of concrete people who are embodied in their social and natural worlds as much as first-person accounts... (p.9).

In other words, the relevant phenomenological reductions aim at irreducible structures of not only the act of consciousness but of the objects of consciousness as well. Depraz uses the example of the experience of red (p.101). After performing the reductions one is left with the essential structure of an experience of red--presumably, one that is much richer than the one had previous to the reduction. In other words, one gets essential (or the essence of) red (as an object of consciousness), and this is DIFFERENT (and epistemically privileged) from contingent red (red as an object of consciousness before the reductions are performed).

The result seems to be that, after one performs these phenomenological reductions, one is left with a more robust and fuller phenomenal world. The world looks different, feels different. And--here's

the meat of it-- if a neuroscientist performs these reductions, the BRAIN will look and feel different, perhaps quite different. The essential structure of an experience of the brain might be altogether different from the one we have from habitual observation. For example, the way neurons look and seem to behave before the reductions might radically differ from the way they look and seem to behave after--if what we get after can even be called a neuron any longer (i.e., things might look VERY different!). With a more robust phenomenal world we would presumably get a more robust neuroscience (since most of neuroscientific data is predicated on observation). After some progress on the first-person side of the project, it would seem necessary that, with this honed observational tool, we revamp the third-person side. Only then could we be said to be on our way to a RIGOROUS science of consciousness (after all, this was Husserl's goal).

Of course, this is only a criticism of someone interested in a science of consciousness who takes Husserlian phenomenology seriously as a means of gathering first-person data (specifically, someone like Varela). If one doesn't take Husserl seriously, one doesn't have to take the criticism seriously. As someone who does take Husserl seriously, it seems that, given the above, it is from the first-person perspective that a science of consciousness must begin. The task of gathering third-person data follows only after significant work has been done on the first-person side. Work done in the third-person beforehand, is, if Husserl is right, haphazard. It may be science, but not, as Husserl says, rigorous science.

From acboch@U.Arizona.EDU Wed Sep 29 20:04:40 1999
Return-Path: <acboch@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id UAA24548
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 29 Sep 1999 20:04:40 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
id <01JGJZKGPW5SB8M4GQ@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Wed, 29 Sep 1999 20:04:54 MST
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU)
by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
with ESMTP id <01JGJZKFRP68B8WQAN@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Wed, 29 Sep 1999 20:04:53 -0700 (MST)
Received: from localhost (acboch@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id UAA23755; Wed,
29 Sep 1999 20:04:51 -0700 (MST)
Date: Wed, 29 Sep 1999 20:04:51 -0700 (MST)
From: Adam C Boch <acboch@U.Arizona.EDU>
Subject: Re: NCC postings
In-reply-to: <199909290630.XAA22086@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909291732240.3534-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

Chalmers wrote:

>
> ADAM: Adam worries that the whole NCC approach is premature, since we
> need to get a good phenomenological method off the ground first. Not
> just to gather the first-person data, but also to gather the
> third-person data from neuroscience, etc. My worry here is that the
> same presumably goes for third-person data in any science at all, e
> g. physics, chemistry, biology. But if Newton had waited for
> phenomenology to be properly developed, we would have been waiting a
> long time! My own view is that a developed phenomenology is vital for
> gathering first-person data, and may end up helping us gather
> third-person data, but that at least the third-person part of the
> story ought to be able to get off the ground without it. It's not
> clear why the NCC search is any worse off here than any other area of
> neuroscience, or of science in general.

> --Dave.

Exactly. Husserl's methodology was intended as an epistemological overhaul of the sciences--i.e., ALL sciences. What he was looking for was a rigorous method, a reliable means to put it another way, of gathering data for an epistemically privileged science. A contemporary analytic analog might be the theory-ladenness of perception/observation doctrine--which, in short, claims that conceptual learning influences the mode and content of perception. If perception is theory-laden, proponents of the doctrine say, we may pose an epistemological criticism of traditional science--wherein the data that is to be explained is gathered via observation. This is a real worry in the philosophy of science. That is, it has well-respected advocates and the debate is far from being putatively settled. If Husserl (and theory-ladenness advocates) are right, we MUST proceed with the first-person project first. Had we recognized this earlier and directed our focus appropriately, perhaps Newtonian physics would have looked quite different and maybe even gotten things right!

Adam

From lachter@u.arizona.edu Thu Sep 30 10:44:34 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id KAA25454
for <chalmers@paradox.soc-sci.arizona.edu>; Thu, 30 Sep 1999 10:44:34 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGKUBDG028B8W6XN@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Thu, 30 Sep 1999 10:44:50 MST
Received: from trifid.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JGKUBCCMZ4B8WHFV@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Thu, 30 Sep 1999 10:44:47 -0700 (MST)
Received: from [10.0.2.15] ([128.196.99.98])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id KAA23536; Thu,
30 Sep 1999 10:44:45 -0700
Date: Thu, 30 Sep 1999 10:44:59 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: NCC postings
In-reply-to: <Pine.HPX.4.10.9909291732240.3534-1000000@pavo.U.Arizona.EDU>
X-Sender: lachter@pop.u.arizona.edu
To: Adam C Boch <acboch@u.arizona.edu>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu

Message-id: <v04020a05b41942209643@[10.0.2.15]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
References: <199909290630.XAA22086@paradox.soc-sci.arizona.edu>
Status: R0

Okay, maybe I am too hung up with my own traditional psychological way of thinking but I really don't understand this first person data business and I am hoping that someone can explain it to me. It seems to me that if we understand the third person perspective in a reasonably thorough manner we can predict any possible finding from the use of first person data. That is because our beliefs about what we are experiencing are accessible from a third person view point. It is hard to see what good first person evidence is going to do us if we don't believe it, but if we believe it it is accessible as third person evidence.

Joel

At 8:04 PM -0700 9/29/99, Adam C Boch wrote:

>Chalmers wrote:

>

>>

>> ADAM: Adam worries that the whole NCC approach is premature, since we
>> need to get a good phenomenological method off the ground first. Not
>> just to gather the first-person data, but also to gather the
>> third-person data from neuroscience, etc. My worry here is that the
>> same presumably goes for third-person data in any science at all, e
>> g. physics, chemistry, biology. But if Newton had waited for
>> phenomenology to be properly developed, we would have been waiting a
>> long time! My own view is that a developed phenomenology is vital for
>> gathering first-person data, and may end up helping us gather
>> third-person data, but that at least the third-person part of the
>> story ought to be able to get off the ground without it. It's not
>> clear why the NCC search is any worse off here than any other area of
>> neuroscience, or of science in general.

>>

>> --Dave.

>>

>>

>Exactly. Husserl's methodology was intended as an epistemological
>overhaul of the sciences--i.e., ALL sciences. What he was looking for was
>a rigorous method, a reliable means to put it another way, of gathering
>data for an epistemically privileged science. A contemporary analytic
>analog might be the theory-ladenness of perception/observation
>doctrine--which, in short, claims that conceptual learning influences the
>mode and content of perception. If perception is theory-laden, proponents
>of the doctrine say, we may pose an epistemological criticism of
>traditional science--wherein the data that is to be explained is gathered
>via observation. This is a real worry in the philosophy of science.
>That is, it has well-respected advocates and the debate is far being
>putatively settled. If Husserl (and theory-ladenness advocates) are
>right, we MUST proceed with the first-person project first. Had we
>recognized this earlier and directed our focus appropriately, perhaps
>Newtonian physics would have looked quite different and maybe even
>gotten things right!

>

>Adam

From acboch@U.Arizona.EDU Fri Oct 1 08:30:10 1999

Return-Path: <acboch@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id IAA00678
for <chalmers@paradox.soc-sci.arizona.edu>; Fri, 1 Oct 1999 08:30:10 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGMILQONK0B8MGII@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Fri, 1 Oct 1999 15:31:09 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGMILP07ZKB8WR0S@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Fri, 01 Oct 1999 15:31:08 -0700 (MST)
Received: from nevis.u.arizona.edu (IDENT:acboch@shell [128.196.137.19])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id IAA00667 for
<scicon@paradox.soc-sci.arizona.edu>; Fri, 01 Oct 1999 08:29:56 -0700
Received: from localhost (acboch@localhost)
by nevis.u.arizona.edu (8.8.8/8.8.8) with ESMTP id PAA24834; Fri,
01 Oct 1999 15:30:42 -0700
Date: Fri, 01 Oct 1999 15:30:42 -0700 (MST)
From: Adam C Boch <acboch@U.Arizona.EDU>
Subject: Re: NCC postings
In-reply-to: <v04020a05b41942209643@[10.0.2.15]>
To: Joel Lachter <lachter@U.Arizona.EDU>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9910011449350.50112-100000@nevis.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Joel (and all listening in),

I'm having some trouble understanding your post, especially the stuff
about belief in the last two sentences.

You start with:

It seems to me that if we understand the third-person
perspective in a reasonably thorough manner we can predict any possible
finding from the use of first-person data.

You seem to be saying that once we have an adequately robust neuroscience,
we can correlate brain states (or processes) with our first-person data,
and so be able to, from then on, predict first-person experience just from
looking at the third-person data.

But then you say that this is the case because, "...our beliefs about what
we are experiencing are accessible from a third-person viewpoint".

First, I don't understand the role beliefs are (or ought to be) playing
here. Second, you seem to be saying that after we correlate third-person
and first-person data, by looking at the former we can predict the latter,
and that this is BECAUSE first-person experience is accessible from
third-person data. But this doesn't seem to be saying anything
substantial.

Finally, you write:

It is hard to see what good first-person evidence is going to do us if we
don't believe it, but if we believe it it is accessible as third-person
evidence.

Here, again, I'm not sure I understand what role beliefs are playing (or,
perhaps I think you mean one thing by beliefs but you really mean
something else). Also, I'm unsure in what way you are using
'third-person data' and 'first-person data'. At first, I thought you just
meant to point to observations in neuroscience when you uses 'third-person
data'. But then you seemed to mean observation of verbal reports when you

used the phrase later. But verbal reports can also be considered indicators of first-person experience.

I hope that you will take the time to clarify these points. Perhaps just mentally noting them and re-phrasing your point would be enough. I am excited about this dialogue and eagerly await your reply.

Adam

On Thu, 30 Sep 1999, Joel Lachter wrote:

> Okay, maybe I am too hung up with my own traditional psychological way of
> thinking but I really don't understand this first person data business and
> I am hoping that someone can explain it to me. It seems to me that if we
> understand the third person perspective in a reasonably thorough manner we
> can predict any possible finding from the use of first person data. That is
> because our beliefs about what we are experiencing are accessible from a
> third person view point. It is hard to see what good first person evidence
> is going to do us if we don't believe it, but if we believe it it is
> accessible as third person evidence.

>
>
> Joel

>
> At 8:04 PM -0700 9/29/99, Adam C Boch wrote:
> >Chalmers wrote:

> >
> >>
> >> ADAM: Adam worries that the whole NCC approach is premature, since we
> >> need to get a good phenomenological method off the ground first. Not
> >> just to gather the first-person data, but also to gather the
> >> third-person data from neuroscience, etc. My worry here is that the
> >> same presumably goes for third-person data in any science at all, e
> >> g. physics, chemistry, biology. But if Newton had waited for
> >> phenomenology to be properly developed, we would have been waiting a
> >> long time! My own view is that a developed phenomenology is vital for
> >> gathering first-person data, and may end up helping us gather
> >> third-person data, but that at least the third-person part of the
> >> story ought to be able to get off the ground without it. It's not
> >> clear why the NCC search is any worse off here than any other area of
> >> neuroscience, or of science in general.

> >>
> >> --Dave.

> >>
> >>
> >Exactly. Husserl's methodology was intended as an epistemological
> >overhaul of the sciences--i.e., ALL sciences. What he was looking for was
> >a rigorous method, a reliable means to put it another way, of gathering
> >data for an epistemically privileged science. A contemporary analytic
> >analog might be the theory-ladenness of perception/observation
> >doctrine--which, in short, claims that conceptual learning influences the
> >mode and content of perception. If perception is theory-laden, proponents
> >of the doctrine say, we may pose an epistemological criticism of
> >traditional science--wherein the data that is to be explained is gathered
> >via observation. This is a real worry in the philosophy of science.
> >That is, it has well-respected advocates and the debate is far being
> >putatively settled. If Husserl (and theory-ladenness advocates) are
> >right, we MUST proceed with the first-person project first. Had we
> >recognized this earlier and directed our focus appropriately, perhaps

> >Newtonian physics would have looked quite different and maybe even
> >gotten things right!
> >
> >Adam
>
>

From lachter@u.arizona.edu Fri Oct 1 14:41:03 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id OAA00842
for <chalmers@paradox.soc-sci.arizona.edu>; Fri, 1 Oct 1999 14:41:03 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGMVJL0JE8B8VHRG@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Fri, 1 Oct 1999 21:42:02 MST
Received: from trifid.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JGMVJJWFNKB8WUR3@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Fri, 01 Oct 1999 21:42:01 -0700 (MST)
Received: from [128.196.99.98] (tec3.Psych.arizona.edu [128.196.98.11])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id VAA26712; Fri,
01 Oct 1999 21:41:57 -0700
Date: Fri, 01 Oct 1999 21:42:09 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: NCC postings
In-reply-to: <Pine.A41.4.10.9910011449350.50112-100000@nevis.u.arizona.edu>
X-Sender: lachter@pop.u.arizona.edu
To: Adam C Boch <acboch@u.arizona.edu>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a02b41af7b34793@[128.196.99.98]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
References: <v04020a05b41942209643@[10.0.2.15]>
Status: RO

Let me see if I can be clearer.

Step one. I do not see any way of having first person data which the person whose first person data it is does not believe. This is not supposed to be any claim about NCCs or something which we might discover. This seems to me like a conceptual truth. If I had some experience, but I did not believe that I had it I don't see how I could use it as data. Furthermore, if I believe I have a certain experience I will count it as data. Someone might convince me that I am incorrect in thinking that I had this experience, but then I will throw out the data just as I no longer believe I had the experience. Thus it seems to me that the information available from first person data is the same information available from beliefs about experience. I don't think this is an a posteriori discovery. I think this is conceptually necessary: Data is data only insofar as you believe it.

Step two. Beliefs are accessible via third person methods. Unlike qualia, we have third person methods for getting at beliefs. We have every reason to think that we will have a very good understanding of beliefs based on cognitive psychology and neuroscience. Anticipating an objection, it may be that we cannot completely individuate beliefs without qualia. If you and I have inverted qualia, one might argue that there is some part of our color beliefs that is different. However, whatever there might be to beliefs that is not accessible to third person analysis, it is not structural. Whatever theory you develop with your inverted qualia will be structurally identical to one based on my qualia or one simply based on the third person data. Notice that any data which you could potentially communicate (even through

demonstration) is third person data. Thus it is not clear how useful any residual first person properties of beliefs are going to be in any science of consciousness.

Step three. From these it follows all the information available from first person methodologies is also available through third person methodologies (modulo qualia inversion type issues).

Joel

At 3:30 PM -0700 10/1/99, Adam C Boch wrote:

>Joel (and all listening in),

>

>I'm having some trouble understanding your post, especially the stuff

>about belief in the last two sentences.

>

>You start with:

>It seems to me that if we understand the third-person

>perspective in a reasonably thorough manner we can predict any possible

>finding from the use of first-person data.

>

>You seem to be saying that once we have an adequately robust neuroscience,

>we can correlate brain states (or processes) with our first-person data,

>and so be able to, from then on, predict first-person experience just from

>looking at the third-person data.

>

>But then you say that this is the case because, "...our beliefs about what

>we are experiencing are accessible from a third-person viewpoint".

>

>First, I don't understand the role beliefs are (or ought to be) playing

>here. Second, you seem to be saying that after we correlate third-person

>and first-person data, by looking at the former we can predict the latter,

>and that this is BECAUSE first-person experience is accessible from

>third-person data. But this doesn't seem to be saying anything

>substantial.

>

>Finally, you write:

>It is hard to see what good first-person evidence is going to do us if we

>don't believe it, but if we believe it it is accessible as third-person

>evidence.

>

>Here, again, I'm not sure I understand what role beliefs are playing (or,

>perhaps I think you mean one thing by beliefs but you really mean

>something else). Also, I'm unsure in what way you are using

>'third-person data' and 'first-person data'. At first, I thought you just

>meant to point to observations in neuroscience when you uses 'third-person

>data'. But then you seemed to mean observation of verbal reports when you

>used the phrase later. But verbal reports can also be considered

>indicators of first-person experience.

>

>I hope that you will take the time to clarify these points. Perhaps just

>mentally noting them and re-phrasing your point would be enough. I am

>excited about this dialogue and eagerly await your reply.

>

>Adam

>

>

>

>On Thu, 30 Sep 1999, Joel Lachter wrote:

>
>> Okay, maybe I am too hung up with my own traditional psychological way of
>> thinking but I really don't understand this first person data business and
>> I am hoping that someone can explain it to me. It seems to me that if we
>> understand the third person perspective in a reasonably thorough manner we
>> can predict any possible finding from the use of first person data. That is
>> because our beliefs about what we are experiencing are accessible from a
>> third person view point. It is hard to see what good first person evidence
>> is going to do us if we don't believe it, but if we believe it it is
>> accessible as third person evidence.
>>
>>
>> Joel
>>
>>
>> At 8:04 PM -0700 9/29/99, Adam C Boch wrote:
>> >Chalmers wrote:
>> >
>> >>
>> >> ADAM: Adam worries that the whole NCC approach is premature, since we
>> >> need to get a good phenomenological method off the ground first. Not
>> >> just to gather the first-person data, but also to gather the
>> >> third-person data from neuroscience, etc. My worry here is that the
>> >> same presumably goes for third-person data in any science at all, e
>> >> g. physics, chemistry, biology. But if Newton had waited for
>> >> phenomenology to be properly developed, we would have been waiting a
>> >> long time! My own view is that a developed phenomenology is vital for
>> >> gathering first-person data, and may end up helping us gather
>> >> third-person data, but that at least the third-person part of the
>> >> story ought to be able to get off the ground without it. It's not
>> >> clear why the NCC search is any worse off here than any other area of
>> >> neuroscience, or of science in general.
>> >>
>> >> --Dave.
>> >>
>> >>
>> >Exactly. Husserl's methodology was intended as an epistemological
>> >overhaul of the sciences--i.e., ALL sciences. What he was looking for was
>> >a rigorous method, a reliable means to put it another way, of gathering
>> >data for an epistemically privileged science. A contemporary analytic
>> >analog might be the theory-ladenness of perception/observation
>> >doctrine--which, in short, claims that conceptual learning influences the
>> >mode and content of perception. If perception is theory-laden, proponents
>> >of the doctrine say, we may pose an epistemological criticism of
>> >traditional science--wherein the data that is to be explained is gathered
>> >via observation. This is a real worry in the philosophy of science.
>> >That is, it has well-respected advocates and the debate is far being
>> >putatively settled. If Husserl (and theory-ladenness advocates) are
>> >right, we MUST proceed with the first-person project first. Had we
>> >recognized this earlier and directed our focus appropriately, perhaps
>> >Newtonian physics would have looked quite different and maybe even
>> >gotten things right!
>> >
>> >Adam
>>
>>

From chalmers@paradox.soc-sci.arizona.edu Mon Oct 4 23:42:53 1999

Return-Path: <chalmers@arizona.edu>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA06959

for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 4 Oct 1999 23:42:52 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JGR6NWTQS0B8X97J@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Mon, 4 Oct 1999 23:43:16 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU

(PMDF V5.1-12 #24137) with ESMTP id <01JGR6NVRXTCB8XEJK@Telcom.Arizona.EDU>

for chalmers@Arizona.EDU; Mon, 04 Oct 1999 23:43:14 -0700 (MST)

Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu

(8.9.3/8.9.3) id XAA06950 for scicon; Mon, 04 Oct 1999 23:42:31 -0700

Date: Mon, 04 Oct 1999 23:42:31 -0700

From: David Chalmers <chalmers@arizona.edu>

Subject: First-person data

To: scicon@paradox.soc-sci.arizona.edu

Message-id: <199910050642.XAA06950@paradox.soc-sci.arizona.edu>

MIME-version: 1.0

Content-type: TEXT/PLAIN; CHARSET=US-ASCII

Status: R

Hi all, I'm back from Memphis. Here are some thoughts on recent discussion topics. First, the issue of reducing first-person data to third-person data.

I think there is definitely something important in what Joel says. Given that first-person data are only usable if they give rise to beliefs, and given that beliefs always show up in behavioral dispositions such as reportability, then it may seem that first-person data insofar as they are usable can be found in third-person data. A few comments that might be made in response:

(i) Even if the content of first-person data had analogs in third-person data (as is suggested e.g. by my principle of structural coherence), this wouldn't mean that the data themselves are eliminable or reducible. It would just mean that we have a natural mapping between the two. Of course this mapping is used all the time by experimenters who use reportability etc as their epistemic route to consciousness. Note that the reports and beliefs themselves aren't so much what the experimenters care about (as they would be if they were studying third-person data); rather, they are mostly interested in them as a guide to first-person phenomenal experience.

(ii) It could be that there are first-person data that don't give rise to beliefs, or that can't be articulated in reports. If so, they may be usable in our science, but that doesn't mean their reality as data must be denied.

(iii) There may be aspects of the first-person data that can only be understood from a first-person point of view. The obvious example is Joel's inverted spectrum case. For someone who has experienced color, the distinction between red and blue first-person data is obvious, but it doesn't really show up in the third-person data. Does that mean we can't have a science of it? Well, if we can assume that our scientists have the concepts of red and blue experience, then we can, as long as our science includes interpreted first-person concepts such as these. Whereas from the third-person case, we'd have to give up right away. Of course, this method isn't useful for those without the concepts, such as Mary in her room. Still, it can take us some distance (and e.g. suggests that Robinson Crusoe might be able to come up with a rich first-person science that third-person data wouldn't approach).

There are also more general versions of this problem (not just for

"intrinsic" inverted spectrum cases), e.g. in understanding the very concept of consciousness and first-person data. Someone without consciousness couldn't even start on this part of the science, for all their grasp of the third-person data, since they wouldn't know what we're talking about re first-person data. But a conscious scientist has no problems at all. So one shouldn't underestimate the role of first-person concept possession in individuating the subject matter of our science here. And it may be the possession of those concepts that makes third-person data indirectly usable in just the way Joel suggests. Without it, one couldn't use third-person data as a guide to first-person data. One would be stuck with zombie science.

--Dave.

From lachter@u.arizona.edu Tue Oct 5 08:13:40 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id IAA07423
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 5 Oct 1999 08:13:40 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGROI8983KB8VXNW@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 5 Oct 1999 08:14:05 MST
Received: from trifid.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JGROI6WOZKB8WV30@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Tue, 05 Oct 1999 08:14:03 -0700 (MST)
Received: from [10.0.2.15] (tec3.Psych.arizona.edu [128.196.98.11])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id IAA26704; Tue,
05 Oct 1999 08:14:00 -0700
Date: Tue, 05 Oct 1999 08:14:17 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: First-person data
In-reply-to: <199910050642.XAA06950@paradox.soc-sci.arizona.edu>
X-Sender: lachter@pop.u.arizona.edu
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a00b41fc3532ece@[128.196.99.98]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
Status: R

At 11:42 PM -0700 10/4/99, David Chalmers wrote:
> One would be stuck with zombie science.

Exactly my point. A zombie (being exactly the same as a normal person physically) could be just as good a visual scientist as anyone else. (i) A zombie would ask to see the stimuli just like anyone else. It is not like he is just faking it when he asks. He wants to see the data that everyone else sees. And he does. Seeing the stimuli causes him to have beliefs about the strategies used by the subjects in the experiment and the information provided by the stimuli just like anyone else. (ii) It seems to me that any "first person data" that can be used when collected from a non-zombie, can also be collected from a zombie. But in what sense is it "first person" then? (iii) The zombie distinguishes red and blue. The zombie can argue about the inverted spectrum problem.

What is it about our science of vision that we could not discover if we were zombies?

Joel

>Hi all, I'm back from Memphis. Here are some thoughts on recent
>discussion topics. First, the issue of reducing first-person data
>to third-person data.
>
>I think there is definitely something important in what Joel says.
>Given that first-person data are only usable if they give rise to
>beliefs, and given that beliefs always show up in behavioral
>dispositions such as reportability, then it may seem that first-person
>data insofar as they are usable can be found in third-person data. A
>few comments that might be made in response:
>
>(i) Even if the content of first-person data had analogs in
>third-person data (as is suggested e.g. by my principle of structural
>coherence), this wouldn't mean that the data themselves are eliminable
>or reducible. It would just mean that we have a natural mapping
>between the two. Of course this mapping is used all the time by
>experimenters who use reportability etc as their epistemic route to
>consciousness. Note that the reports and beliefs themselves aren't so
>much what the experimenters care about (as they would be if they were
>studying third-person data); rather, they are mostly interested in
>them as a guide to first-person phenomenal experience.
>
>(ii) It could be that there are first-person data that don't give rise
>to beliefs, or that can't be articulated in reports. If so, they may
>be usable in our science, but that doesn't mean their reality as data
>must be denied.
>
>(iii) There may be aspects of the first-person data that can only be
>understood from a first-person point of view. The obvious example is
>Joel's inverted spectrum case. For someone who has experienced color,
>the distinction between red and blue first-person data is obvious, but
>it doesn't really show up in the third-person data. Does that mean we
>can't have a science of it? Well, if we can assume that our
>scientists have the concepts of red and blue experience, then we can,
>as long as our science includes interpreted first-person concepts such
>as these. Whereas from the third-person case, we'd have to give up
>right away. Of course, this method isn't useful for those without the
>concepts, such as Mary in her room. Still, it can take us some
>distance (and e.g. suggests that Robinson Crusoe might be able to come
>up with a rich first-person science that third-person data wouldn't
>approach).
>
>There are also more general versions of this problem (not just for
>"intrinsic" inverted spectrum cases), e.g. in understanding the very
>concept of consciousness and first-person data. Someone without
>consciousness couldn't even start on this part of the science, for all
>their grasp of the third-person data, since they wouldn't know what
>we're talking about re first-person data. But a conscious scientist
>has no problems at all. So one shouldn't underestimate the role of
>first-person concept possession in individuating the subject matter of
>our science here. And it may be the possession of those concepts that
>makes third-person data indirectly usable in just the way Joel
>suggests. Without it, one couldn't use third-person data as a guide
>to first-person data. One would be stuck with zombie science.
>
>--Dave.

From lachter@u.arizona.edu Tue Oct 5 08:25:43 1999

Return-Path: <lachter@u.arizona.edu>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id IAA07439
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 5 Oct 1999 08:25:43 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGROX6B6E8B8XKLU@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 5 Oct 1999 08:26:08 MST
Received: from trifid.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTTP id <01JGROX4UJE0B8XGW6@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Tue, 05 Oct 1999 08:26:06 -0700 (MST)
Received: from [10.0.2.15] (tec3.Psych.arizona.edu [128.196.98.11])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTTP id IAA27888; Tue,
05 Oct 1999 08:26:03 -0700

Date: Tue, 05 Oct 1999 08:26:20 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Oops...(Re: First-person data)
X-Sender: lachter@pop.u.arizona.edu
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a01b41fc7e44292@[10.0.2.15]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
Status: R

At 8:19 AM -0700 10/5/99, Joel Lachter wrote:
> One would be stuck with zombie science.

>visual scientist
>science of vision
...

I don't think it matters if you substitute consciousness scientist, science
of consciousness etc (or anything else for that matter). But since it is a
science of consciousness not vision that is at issue, I should have used
"consciousness" not "vision."

Joel

From acboch@U.Arizona.EDU Tue Oct 5 14:33:04 1999
Return-Path: <acboch@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id OAA01439
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 5 Oct 1999 14:33:04 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGS1Q59DQ8B8XBQA@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 5 Oct 1999 14:33:05 MST
Received: from nevis.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTTP id <01JGS1Q3X6U0B8X55A@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Tue, 05 Oct 1999 14:33:04 -0700 (MST)
Received: from localhost (acboch@localhost)
by nevis.u.arizona.edu (8.8.8/8.8.8) with ESMTTP id OAA35332; Tue,
05 Oct 1999 14:33:01 -0700
Date: Tue, 05 Oct 1999 14:33:01 -0700 (MST)
From: Adam C Boch <acboch@U.Arizona.EDU>
Subject: Re: NCC postings
In-reply-to: <v04020a02b41af7b34793@[128.196.99.98]>
To: Joel Lachter <lachter@U.Arizona.EDU>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9910051314410.27358-100000@nevis.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R0

Joel,

I guess the source of my confusion was the following. I was trying to interpret your point regarding beliefs about first-person data vis-a-vis my point regarding Husserlian reductions.

That was a mistake since it seems to be a separate point.

But now I would like to address the point by invoking some of the notions I brought out for discussion before.

You wanted to say that all this business about first-person experience was superfluous, since any relevant first-person data (i.e., the ones a subject has beliefs about) can be had by third-person investigation, viz., observation of subject reports.

But, there are still problems (Dave mentioned a few. I'll mention some more, maybe overlapping a bit with his stuff.).

There might be much we are missing if we limit first-person data to those that we can glean from subject reports.

(1) There may be important experiential data that subjects are not forming beliefs about.

I talked before about possible "presentations" of objects of consciousness. These are other 'points of view' to put it simply, other perspectives with respect to objects of consciousness and modes of being conscious. If Husserl is right, in virtue of relying solely on our habitual perspectives we miss out on the other possibilities. That is, we are conscious of less than is possible and so limit our domain of inquiry. If our domain of conscious experience is restricted, so is our domain of belief with respect to what we are conscious of.

(2) There are some experiences that just cannot adequately be reported.

Take as an example, G.Y., a subject used in the blindsight experiments we've discussed. The experimenters made a point to convey statements made by G.Y. to the effect that he could not accurately explain to the experimenters the content of his "blindsight" experiences. He would say things like "the closest I could come to describing it", which points to his inability to EXACTLY describe it. Other subjects in blindsight experiments express similar difficulties. The notion seems to be that 'if you haven't experienced it for yourself, you can't know what I'm talking about'.

Luckily, a phenomenologically-based research project eliminates these problems.

Regarding (1), those other possible perspectives of objects (or modes) of consciousness are opened up via employment of the reductions. One's domain of conscious experience expands thereby expanding ones domain of belief with respect to conscious experience.

But the point about belief and reportability is moot since...

Regarding (2), in a phenomenologically-based research program, each individual experimenter is herself her own subject. Each individual investigator performs his own reductions. When a given phenomenon is experienced by all, a common language to describe it is set up, and intersubjectivity is reached.

Using blindsight experimentation as an example, we can see the limitations

pointed to above. The subjects, by the very nature of their maladies, are limited with respect to their domain of P-consciousness. Also, they have trouble describing to their investigators just what they ARE P-conscious of.

Perhaps, using a different methodological paradigm, namely a Husserlian phenomenological one, we might be able to experience the 'blindsight' phenomenon for ourselves. That is, the phenomenon would be one amongst a variety of different modes of being conscious. Each investigator could perform the necessary reductions such that she came to have the relevant experience. Perhaps then, other, closely related modes could be experienced, and the data compared with that of other investigators (who themselves had had the same experiences), to arrive at some objective (i.e., intersubjective) explanation of the matter (of course this might need to include the relevant 'neural' data, but, as I have said before, this itself must be done only after the relevant reductions have been performed (viz., those having to do with physical objects, motion, tie, etc., i.e., the one's pertinent to doing brain science)).

Perhaps notions of A and P consciousness would become more clearly individuated (or perhaps they would bleed into each other).

Regardless of the outcome, the point is that the methodological tools provided by phenomenology provide a solution to problems that arise upon employment of more traditional methodological tools, namely, those of experimental psychology.

Cheers,

Adam

From chalmers@paradox.soc-sci.arizona.edu Thu Sep 23 18:13:31 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id SAA10855
for <chalmers@paradox.soc-sci.arizona.edu>; Thu, 23 Sep 1999 18:13:31 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGBHXDFZFKB8VN0F@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Thu, 23 Sep 1999 18:13:35 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGBHXA9U9SB8V4JG@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Thu, 23 Sep 1999 18:13:31 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id SAA10848 for scicon; Thu, 23 Sep 1999 18:13:22 -0700
Date: Thu, 23 Sep 1999 18:13:22 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: What does a neural correlate of consciousness explain
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909240113.SAA10848@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=ISO-8859-1
Content-transfer-encoding: 8BIT
Status: R

>From Kniels1@aol.com Thu Sep 23 15:57:22 1999
From: Kniels1@aol.com
Date: Thu, 23 Sep 1999 18:57:14 -0400 (EDT)
Subject: What does a neural correlate of consciousness explain?
To: Chalmers@Arizona.EDU
Cc: lnielsen@u.arizona.edu
X-MIME-Autoconverted: from quoted-printable to 8bit by paradox.soc-sci.arizona.edu id PAA10743

Hi Dave,

I'm out of town and don't have access to my own e-mail accounts, so I can't
post to the list. Would you please submit it for me?

Thanks,

Lis

In his talk of September 10, Dave accorded a prominent place to the search
for the NCC in an overall science of consciousness. He viewed it as playing
an important role in leading to the discovery of bridging principles linking
the physical and phenomenal facts. My comments here address why I think this
expectation is premature, and why approaches like those of Logothetis and
Crick & Koch, while clever and interesting, are unlikely to give us much new
information about consciousness. (I make no claims regarding the originality
of these ideas, as they are greatly influenced by comments made by others in
class. Writing them out has merely helped to focus my thinking on these
issues.)

Whenever we have found specific brain correlates for particular psychological
functions, we seem to feel that we have better understood those functions.
In fact, often all that we have understood is why damage to those brain areas
leads to deficits of this or that sort. In the current climate where
high-tech brain mapping is so fervently pursued, this feeling of making
explanatory progress is reinforced by numerous social factors that drive
research. I remain skeptical that any substantial explanatory progress is
actually made in this field, beyond the continued confirmation and refinement
of our existing hypotheses about brain localization. The following (recently

encountered) quote on the relevance of neural correlates to the shaping of developmental theory expresses part of my worry as regards how this kind of evidence can inform a theory of consciousness:

"Suppose we have a psychological theory that says that function A leads ontogenetically into function B. Being able to suggest brain correlates for A and B has great merits, but it does not by itself advance developmental theory. This can only be done if there is information about those brain correlates that can be used to inform the developmental theory, such as the relationship between physiological substrates of A and B. ... [P]hysiology on its own has never been a prerequisite to developmental theory, since sometimes the psychological and neurological facts match on levels and sometimes they do not." (Segalowitz, 1994):

What is said here in terms of developmental theory seems to apply just as well to discussions of a theory of consciousness. Suppose we have an area IT that we know codes for a certain type of visual information X, because cells in IT fire when subjects are viewing X. We then discover that firing in IT correlates with phenomenal experience of X under conditions of binocular rivalry. We can assume that this is either because some top-down process has focused attention on X (as opposed to the rivalrous stimulus Y) or because some bottom-up process has caused the processing of X to override the processing of Y. What does this tell us about consciousness that we didn't know before? Arguably little, unless "there is information about those brain correlates that can be used to inform the ... theory [of consciousness]." For this to be the case, we need to learn something new about the role of IT in relation to these other aspects of visual processing. For example, something new about its position in the visual processing chain or in informational feedback loops, or something new about its connectivity to other brain areas that sheds light on how these top-down or bottom-up processes operate. Presumably, this is the kind of thing we would like to learn from our neural investigations, namely how the information that (we already know that) the brain encodes is made accessible to consciousness. We want to know WHAT makes them accessible, and what this consciousness IS to which they are accessible. Only insofar as a discovery about the NCC of conscious content can shed light on these issues does it inform our theory in any new sense.

Crick and Koch place certain connectivity demands on any potential NCC (that it have connections to the prefrontal cortex), somehow putting the cart before the horse in this research program. The problem with their approach is that once we assume that we understand the functional connectivity of any NCC, then we are unable to learn anything new about the nature of consciousness from finding its neural correlates. Their research program leaves us with two options. Either we discover that all areas connected with PFC are NCCs for different contents of consciousness. Or we discover that only a subset of these areas are NCCs and have to explain what makes them different from the remainder. We are still left without a clue as to how PFC connections generate consciousness. Until we have an answer to that question, we are simply engaged in more brain mapping.

The approaches of both Logothetis and Crick and Koch make further assumptions about the NCC that are not necessarily justified. First, they assume that phenomenal consciousness and neural firings will correlate on temporal dimensions. But, as Amanda pointed out, they have established no guidelines for what degree of correlation is acceptable. One might be tempted to conclude that binocular rivalry studies confirm the temporal correlation hypothesis, but in fact all they show is that there is a place in the processing stream where there is a (rough) temporal correlation between phenomenal consciousness and visual information processing.

It could be that the neural correlate of consciousness resides in the

circuits independent of content-related processing that are, perhaps, firing all the time, i.e., those circuits that show no deviation from their baseline when we are conscious of a stimulus. That's the activity that usually gets subtracted out of any brain imaging or mapping study - the activity that is constant across all trials. Content-related firings (such as those in IT) might just feed into whatever background circuits are maintaining this overall conscious state, accounting for their differential localization as a function of stimulation or attention. But this is not surprising, once one already knows that the brain compartmentalizes its content in various ways. We had a good idea about this already, from numerous lesion and experimental studies that continue to be confirmed and refined by brain mapping techniques.

What neural activity or circuitry determines whether and how a particular content gets into consciousness seems to me a far more interesting question than what neural areas are active when that particular content is represented. The more piecemeal approaches of the binocular rivalry type are likely to leave us with a huge and unwieldy catalogue of NCCs in multiple modalities and at multiple levels of processing. My guess is that they will tell us little NEW about consciousness. In my opinion, our search for the neural correlates of consciousness should focus on questions of access. They should be guided by our current best understanding of the function of consciousness, and of why certain contents are made available to consciousness at any given time. Finally, until we have a more fine-grained mapping of the phenomenal realm, our questions about what phenomenal content states map onto what neural states would seem premature.

From switanek@U.Arizona.EDU Thu Sep 23 20:32:37 1999
Return-Path: <switanek@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id UAA10960
for <chalmers@paradox.soc-sci.arizona.edu>; Thu, 23 Sep 1999 20:32:37 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
id <01JGBMSTUZHCB8VLW8@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Thu, 23 Sep 1999 20:32:41 MST
Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
with ESMTTP id <01JGBMSRKIWGB8VGV0@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Thu, 23 Sep 1999 20:32:38 -0700 (MST)
Received: from localhost (switanek@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTTP id UAA04202; Thu,
23 Sep 1999 20:32:36 -0700 (MST)
Date: Thu, 23 Sep 1999 20:32:36 -0700 (MST)
From: Nicholas J Switanek <switanek@U.Arizona.EDU>
In-reply-to: <199909240113.SAA10848@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909231850270.25708-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

I'd like to know whether I'm thinking of NCCs in the right way. It appears that an NCC is a sequence of neural firings. Then, characterizing any NCC requires the identification of a set of neurons and the order in which they fire. If this is on the right track, then a horde of issues descends upon me.

1. First, a small point: Dave wrote in his last email that the argument for the sufficiency of an NCC N might be damaged if the brain is. There could be, he suggests, activity in N without the conscious state that N

correlates ever arising. He might mean that there is activity in the set of neurons that partially make up an NCC. But there's no reason to expect that firing alone should give rise to the conscious experience N gives rise to, because N only becomes sufficient for the conscious experience when N is fully instantiated, the neurons in the relevant set fire in the precisely ordered way. The strengths of the argument for sufficiency are not harmed in the case Dave suggests.

2. Second, let there be a lesion. Now, I might imagine that the set of neurons relevant to a particular NCC N might be unscathed by this nasty lesion and the consciousness-bestowing, magical, exact sequence of firings in this set, call the set {N}, takes place, but the conscious experience the mind was wont to have does not arise. Intuitively, we might say that lesions leave us woozy; it's difficult having the same experience after undergoing them. What this situation suggests to me is that there might be important relational properties adhering the neurons of {N}. How the neurons are related to one another and to other parts of the brain--how off(or off-off) Cartesian Broadway they are--might give clues to why they and not another set of neurons are used to instantiate N. Presumably we couldn't string together the elements of {N}(in a string that might cross itself), watch them firecrack in sequence, and expect the conscious experience to arise. The geometry of the set is relevant. Indeed, this idea might suggest an explanation how the damaged brain finds another set of neurons, {M}, in which to allow a certain conscious experience to take place, one that N used to correlate. The relational properties of the neurons in {M} might be required to be isomorphic to those of the elements of {N}, for M and N to be able to give rise to the same conscious experience.

3. Last, I think Lis is right, if I understand her correctly. Apparently, talk of neural correlates just means the brain mapping project is getting closer to the neural level of description and not necessarily that we are any closer to an explanation of consciousness. Maybe sets of neurons that might serve as NCC-platforms might be isolated and identified. Maybe isomorphic functions between neural sets might be described, to explain the flexibility of NCCs. Perhaps rules will be found that govern the interaction of neurons participating in an NCC, rules that might be strongly correlated to certain types of conscious experience. Won't the hard question remain?

A last, interesting paragraph of Lis's is on the extension of experience and its neural correlate through time. It calls to my mind arguments about the nature of time. We think ourselves to be instantaneously conscious, but clearly we wouldn't say that the neurons firing in an NCC at a particular instant give rise to that instant's experience.

Nick

ps Sorry for all the wanton speculation. I hope the analogues weren't too loose. I am grateful to any who would explain me my folly.

From hvorecky@U.Arizona.EDU Fri Sep 24 11:51:43 1999

Return-Path: <hvorecky@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id LAA12164

for <chalmers@paradox.soc-sci.arizona.edu>; Fri, 24 Sep 1999 11:51:43 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JGCIWB49EOA62REM@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Fri, 24 Sep 1999 11:51:45 MST

Received: from f1n2.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

with ESMTTP id <01JGCIW82KBKB8VIVJ@Telcom.Arizona.EDU> for

chalmers@Arizona.EDU; Fri, 24 Sep 1999 11:51:41 -0700 (MST)

Received: from localhost (hvorecky@localhost)
by f1n2.u.arizona.edu (8.8.8/8.8.8) with ESMTP id LAA20926; Fri,
24 Sep 1999 11:51:40 -0700
Date: Fri, 24 Sep 1999 11:51:40 -0700 (MST)
From: Juraj Hvorecky <hvorecky@U.Arizona.EDU>
Subject: Neural correlate of consciousness
In-reply-to: <199909240113.SAA10848@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909241142090.23108-100000@f1n2.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

As I was thinking about the problem David rises concerning the issues of lesion and their relation to discrimination of the NCC, it now looks to me that if his reasoning is correct (and I believe it is), the distinction between "good" and "bad" lesions is no more than necessary. Why? Well, because it not only applies to the search for NCC, but to any neurological correlate of any disorder. Take any dissociation studies, such as prosopagnosia.

In order to study it, neuroscientists have determined a possible "prosopagnosia correlate", that is a part of brain, which, if damaged, patients are unable to recognize familiar faces. But what if the reasoning is mistaken and the thing they have identified is only somehow part of a causal chain, which leads to face recognition. That is, the "real" correlate is more downstream, but this particular place is so important, that it in fact bears all the important signs of the real correlate. Does it make sense?

Well, if it does, then we have to face consequences. One of them would be, that the proposals Dave's talk about as correct ways to determine correlates (unusual percepts and direct brain stimulation) might not also be so helpful. How do you, for example, want to test "correlate" of faces recognition with UNUSUAL stimulus? What is an unusual face. If the correlate functions only in usual conditions, then such cases might not work.

We would still be safe as far as the brain stimulation works, but concerning the overall technical difficulties, the sceptic might conclude that the time we determine the real correlate is infinitely far away...

Great weekend to everyone!

From logant@U.Arizona.EDU Sat Sep 25 13:48:13 1999
Return-Path: <logant@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id NAA13855
for <chalmers@paradox.soc-sci.arizona.edu>; Sat, 25 Sep 1999 13:48:13 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGE1974Q7KB8VWPU@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sat, 25 Sep 1999 13:48:20 MST
Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JGE195EN80B8VT5T@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Sat, 25 Sep 1999 13:48:18 -0700 (MST)
Received: from localhost (logant@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id NAA14264; Sat,
25 Sep 1999 13:48:16 -0700 (MST)
Date: Sat, 25 Sep 1999 13:48:16 -0700 (MST)
From: Logan T Trujillo <logant@U.Arizona.EDU>
Subject: Re: What does a neural correlate of consciousness explain
In-reply-to: <199909240113.SAA10848@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>

Cc: scicon@paradox.soc-sci.arizona.edu

Message-id: <Pine.HPX.4.10.9909251308590.5985-100000@orion.U.Arizona.EDU>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: RO

Hello All:

I have a couple of comments to make concerning the discussion of NCCs. First, Lis Nielsen recently made a post that I believe is relevant to the arguments Brad and I made on Tuesday concerning the difficulty in localizing a NCC. Here is an excerpt from her posting:

> The approaches of both Logothetis and Crick and Koch make further assumptions
> about the NCC that are not necessarily justified. First, they assume that
> phenomenal consciousness and neural firings will correlate on temporal
> dimensions. But, as Amanda pointed out, they have established no guidelines
> for what degree of correlation is acceptable. One might be tempted to
> conclude that binocular rivalry studies confirm the temporal correlation
> hypothesis, but in fact all they show is that there is a place in the
> processing stream where there is a (rough) temporal correlation between
> phenomenal consciousness and visual information processing.
>
> It could be that the neural correlate of consciousness resides in the
> circuits independent of content-related processing that are, perhaps, firing
> all the time, i.e., those circuits that show no deviation from their baseline
> when we are conscious of a stimulus. That's the activity that usually gets
> subtracted out of any brain imaging or mapping study - the activity that is
> constant across all trials. Content-related firings (such as those in IT)
> might just feed into whatever background circuits are maintaining this
> overall conscious state, accounting for their differential localization as a
> function of stimulation or attention. But this is not surprising, once one
> already knows that the brain compartmentalizes its content in various ways.
> We had a good idea about this already, from numerous lesion and experimental
> studies that continue to be confirmed and refined by brain mapping
> techniques.

Lis' comment seems to accord with the metaphor I put forward about IT being the sculptor that acts upon the clay (lower visual cortical areas, i.e. V1-5 et al) to produce a conscious visual percept. What reason do we have to think that visual qualia are restricted to only one area of the visual pathway and not another? The example of blindsight may seem to suggest that certain portions of the visual stream have qualia producing aspects in that damage to these areas leads to a loss of phenomenology on the part of the blindsight patient. However there are two things to consider: (1) if we remove processing at lower levels (as in blindsight) there is no lower level visual information for higher level areas to act upon to create a visual percept (i.e. no clay). Thus both lower and higher visual areas are needed for the creation of a conscious percept. Given this, does it make sense to say that one area is a NCC and the other is not?; (2) as we will see in the case of blindsight, damage to lower visual areas does not preclude the possibility that visual information can still reach conscious awareness. It is just that the phenomenological manifestation of this information within the blindsighted individual is extremely impoverished. Depending upon the extent of the damage, the phenomenology may range from absolutely nothing to vague (almost indescribable) sensations. In the case of the latter, could it be that the vague sensations are produced from higher order visual areas that receive the information via roundabout loops that bypass the lower visual areas? Since this information is not coming through the visual stream

proper, activation of the higher areas may lead to the production of qualia that reflects the higher level activity. Since the lower areas are damaged (and not producing qualia) this "higher level" qualia would not be bound to the "lower level" qualia, and thus the resulting sensations would be "vague" or "nebulous". In light of the fact that we still do not know how neurons produce qualia at all, views that ascribe qualia production to only limited areas of the brain are extremely premature.

My final comment is the following: I think the difficulty with the definition of an NCC that we have been discussing (and it seems to be the core of Lis' argument above) is in the notion of "correlation". Weak correlations and strong correlations are both correlations; what reason do we have to believe a priori that only strong correlations qualify as the proper criteria for a NCC?

Logan T.

From franzen@U.Arizona.EDU Sat Sep 25 17:45:41 1999
Return-Path: <franzen@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA14211
for <chalmers@paradox.soc-sci.arizona.edu>; Sat, 25 Sep 1999 17:45:41 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGE9JMG140B8W6S1@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sat, 25 Sep 1999 17:45:48 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGE9JKXN0WB8V0SV@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sat, 25 Sep 1999 17:45:46 -0700 (MST)
Received: from f1n7.u.arizona.edu
(IDENT:franzen@f1n7.U.Arizona.EDU [128.196.137.107])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA14205 for
<scicon@paradox.soc-sci.arizona.edu>; Sat, 25 Sep 1999 17:44:09 -0700
Received: from localhost (franzen@localhost)
by f1n7.u.arizona.edu (8.8.8/8.8.8) with ESMTP id RAA31970 for
<scicon@paradox.soc-sci.arizona.edu>; Sat, 25 Sep 1999 17:44:18 -0700
Date: Sat, 25 Sep 1999 17:44:18 -0700 (MST)
From: Peter L Franzen <franzen@U.Arizona.EDU>
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909251712380.34708-100000@f1n7.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

It is concluded from the empirical work of Logothetis on vision in monkeys that "V1 is unlikely to be or involve an NCC, for example, due to the failure of V1 cells to correlate with the contents of consciousness." I don't follow the reasoning here. What is being lost is the issue of networks in the brain, and that these networks, through interaction, may bring about the consciousness phenomenon. And, how exactly can the inferior temporal (IT) cortex be the NCC, and not V1, when the IT receives projections, either directly or indirectly, from V1? With all the rush toward localization, do we risk losing ground in the search for NCC by missing the bigger picture? A person can lose a part of their brain that is thought to be important in some function through lesions, but not lose that function completely. There are often multiple pathways, and losing part of the circuit does not guarantee that there are not other ways the circuit can be completed. Of course, it is always possible that there is more than one NCC, and that a lesion in one does not disrupt the other. Or perhaps there is remapping of a function elsewhere in the brain. How would that change a NCC?

As for whether the NCC state is required to be necessity/sufficient for the conscious state... If an NCC state is only required to be sufficient for the corresponding state, but not necessary, are we any further in understanding how the NCC leads to consciousness or an experience of consciousness state? If it is only sufficient, and not necessary, than how do we know its truly important for the conscious state?

From press@U.Arizona.EDU Sun Sep 26 16:08:57 1999
Return-Path: <press@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA16864
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 16:08:57 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGFKH2GMDSB8W1WI@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 26 Sep 1999 16:09:06 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGFKH1DY34B8VWRV@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 26 Sep 1999 16:09:05 -0700 (MST)
Received: from f1n2.u.arizona.edu
(IDENT:press@f1n2.U.Arizona.EDU [128.196.137.102])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA16859 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 16:08:51 -0700
Received: from localhost (press@localhost) by f1n2.u.arizona.edu (8.8.8/8.8.8)
with ESMTP id QAA20664 for <scicon@paradox.soc-sci.arizona.edu>; Sun,
26 Sep 1999 16:09:01 -0700
Date: Sun, 26 Sep 1999 16:09:01 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>
Subject: Minimal sufficiency and redundancy
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909261605160.32328-100000@f1n2.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

All -

When Dave endorses minimal sufficiency as the proper criterion for identifying the NCC, he notes that on this criterion it will be possible that there will be more than one NCC. He seems untroubled by this, and in one sense I share that attitude. Not only should we not be surprised if there is no "Cartesian Theater" where all the representations involved in phenomenal consciousness occur together, but we should not be surprised if there are redundant NCCs. Given the prevalence of redundant systems throughout the body (kidneys, lungs, etc.) it seems to make good evolutionary sense, other things being equal, to have multiple systems for consciousness.

In fact, I think there is strong evidence that this is the case which can be found in studies of split brain patients. Im sure that the psychologists among us know much more about these experiments than I do, and perhaps some of the philosophers too, but just in case some dont, heres a brief (and somewhat idealized) synopsis:

When the bundle of nerve fibers (corpus callosum) that connect the right and left cerebral hemispheres is severed, the patients suffer a dissociation phenomenon. Because each hemisphere receives input from and has control over (primarily) the opposite side of the body, it is possible to create situations in which one hemisphere knows something that the other doesnt. For example, if the experimenter were to show the word

"handbag" to one of these patients in such a way that the "hand" part would only be perceived by the left hemisphere and the "bag" part only by the right hemisphere, and if the patient were then asked to point at the object named (from a more or less random collection of objects provided) , the right hand (controlled by the left hemisphere) would point to, say, a mannequin hand, and the left hand would point to, say, a paper bag. The patient would not point out the purse. Many would say that this is evidence that in the split brain patient there are two independent streams of conscious states.

Though the split brain cases are fascinating in themselves, what they suggest about normal cases is more relevant to my point here. Since cutting the corpus callosum has no effect on the structure of the rest of the brain, it seems reasonable to conclude that the structures responsible for consciousness must be duplicated in each hemisphere even in normal patients. Since each of these structures can function on its own in the split brain cases, each would seem to be an NCC. In normal patients, the corpus callosum seems to have an information sharing function, so that the two NCCs end up with the same content. (For the sake of simplicity, assume that there is a single Cartesian-Theater-style NCC in each hemisphere).

Dave doesn't seem to be worried about this, but I think that perhaps he should. The worry I have is that the existence of multiple NCCs may raise a set of questions to which he cannot provide an answer. For example, if every normal person has a right NCC and a left NCC, how could we decide which of the following views is true.

- 1) A normal person's phenomenal consciousness is produced by both NCCs running in parallel.
- 2) A normal person's phenomenal consciousness is produced by his left NCC. The right NCC performs all the same operations as the left NCC but does so without producing any phenomenal properties.
- 3) 2) but with right and left reversed.
- 4) A normal person has two independent sets of phenomenal consciousness (two minds), one produced by each NCC. Since the two run in parallel, the phenomenal contents of each mind is identical, but they are nevertheless separate minds.

The problem would only get worse if we found that there were other redundancies within the hemispheres, or if the Cartesian Theater model turned out to be inaccurate in such a way that redundant versions of all sorts of representations of conscious contents were scattered about in the brain.

The reason that I think Dave will have a hard time defending one of the options above is that he has no causal explanation of phenomenal properties, merely a correlation. If we had a theory of consciousness that told us that the NCC causes consciousness by emitting "phenomenal particles" which we could detect with a "phenomometer," we might be able to verify that, say, 2) was the correct option because normal subjects only emit "phenomenal particles" from their left hemispheres. But, of course, this is precisely what Dave doesn't think we will ever be able to do. I suppose he could appeal to criteria of simplicity, appeal to the best explanation, and so on, but it is not clear to me that any one of these possible claims is any simpler, or a better explanation, etc. Of course, maybe a theory of consciousness that narrows things down to one of the above alternatives is the best we can do given our epistemological situation, but if so, I think a critic could justifiably claim that our

"explanation" of consciousness is much too weak. After all, if we cannot even tell how many minds (streams of consciousness, conscious states) a person has, how much do we really know about the phenomena we claim to have explained?

Joel

From serobert@U.Arizona.EDU Sun Sep 26 18:33:00 1999
Return-Path: <serobert@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id SAA17248
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 18:32:59 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGFPHTK00B8VRQR@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 26 Sep 1999 18:33:08 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGFPHM3E3KB8VZ5E@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 26 Sep 1999 18:33:06 -0700 (MST)
Received: from f1n3.u.arizona.edu
(IDENT:serobert@f1n3.U.Arizona.EDU [128.196.137.103])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id SAA17243 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 18:32:38 -0700
Received: from localhost (serobert@localhost)
by f1n3.u.arizona.edu (8.8.8/8.8.8) with ESMTP id SAA32280 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 18:32:49 -0700
Date: Sun, 26 Sep 1999 18:32:48 -0700 (MST)
From: Simon E Roberts-Thomson <serobert@U.Arizona.EDU>
Subject: Correlation and Lesion studies
In-reply-to: <Pine.A41.4.10.9909261605160.32328-100000@f1n2.u.arizona.edu>
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909261718330.33452-100000@f1n3.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Dear all,

Many of the recent mails have tended to be rather scathing of either the use of correlation in the search for NCC, or the use (or interpretation) of lesion studies. I thought that it might be interesting to try and defend both of these - perhaps it will help us to focus more clearly on their problems.

(1) Correlation

The criticisms of correlation seem to be centred on the idea that we are failing to advance our knowledge of *consciousness* when we correlate neuronal firing with specific conscious experience. Nicholas seems to suggest that all we are doing is improving our 'maps', and not answering the real questions. In a similar vein, Lis is worried that the *real* work of consciousness might be taking place in an area where firing is unrelated to content, and hence would fail to sow up in these mapping studies. I will try to answer each charge in turn.

With respect to Nicholas' concern, it may well be that all we are doing when we study neuronal firing is improving our 'maps' of the brain. However, it does not seem to me that this is a problem. Presumably any theory of consciousness will need to account for the correlations between specific content and firing patterns - a part of what such studies are doing is defining the things that a brain-based theory of consciousness needs to address. Thus whilst these studies do not in themselves give us a theory, they are essential in that they define part of the problem.

Turning now to Lis' point, it is certainly possible that the real is such that its firing is not specifically related to content. Instead, areas of the brain whose firing is content-related feed into this (continually?-)firing area, which then produces consciousness. Whilst this view is possible, however, it seems to me to be implausible. If the real NCC does not change its firing rate or pattern when confronted with differing stimuli, then how do we have differing content in consciousness? It seems reasonable to assume that if there is a central NCC, then it must use something like differential firing rates to engender different conscious experiences. If this is so, then presumably such changes should appear of the scans (if not, then this must be due to the limits of the equipment, and not the theory behind their use). In other words, there will be some sort of correlation between the firing and the experience.

(2) Lesion Studies.

One of the problems with the use of lesion studies that was discussed both here and in class, was the concern that whilst a particular lesion might result in the loss of a certain type of consciousness, there is no reason to assume that the area in which the lesion is situated is the NCC for that particular type of consciousness. It may be that the region is merely a part of a chain which leads to the particular NCC. Likewise, there might be multiple pathways to an NCC, such that the loss of one does not result in the loss of that particular consciousness. In response to these concerns, I would like to say that this is not a reason not to use lesion studies, but merely a reason to be careful of how such studies are interpreted. Thus if we had one lesion study which suggested that region A was involved in visual consciousness, then we should not necessarily assume that visual consciousness takes place in A, but rather the much weaker claim that A is involved in visual consciousness. If there were to be multiple studies of related lesion studies, all of which seemed to result in a loss of visual consciousness, then perhaps we could 'map out' the visual pathway (of course, if the brain turns out to be very plastic, then this would be more difficult). In a similar way, it might be possible to identify redundant pathways. Thus these studies will help us to identify what parts of the brain are necessary and/or sufficient for consciousness.

I hope that I have managed to put together a passable defence of the above concepts. I think that they are likely to be very important to the study of consciousness, and should not be dismissed lightly.

Simon.

From anhabib@U.Arizona.EDU Sun Sep 26 19:17:36 1999
Return-Path: <anhabib@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id TAA17273
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 19:17:36 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGFR1Z3UW0B8VQA5@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 26 Sep 1999 19:17:46 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGFR1X4GM8B8WMAC@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 26 Sep 1999 19:17:43 -0700 (MST)
Received: from f1n7.u.arizona.edu
(IDENT:anhabib@f1n7.U.Arizona.EDU [128.196.137.107])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id TAA17268 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 19:17:26 -0700
Received: from localhost (anhabib@localhost)

by f1n7.u.arizona.edu (8.8.8/8.8.8) with ESMTP id TAA30020 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 19:17:37 -0700
Date: Sun, 26 Sep 1999 19:17:37 -0700 (MST)
From: Allen N Habib <anhabib@U.Arizona.EDU>
Subject: Some thoughts on the NCC
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9909261829280.30798-100000@f1n7.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Hey all,

I have three thoughts on the NCC that have been bothering me since our class talk. The first is this, why are we limiting our discussion solely to considerations of necessity and sufficiency? If we are willing to admit that there may be more than one NCC for even a very specific conscious state (what Dave would call a content state), and if we are willing (pace Dave) to admit the possibility that an NCC could be instantiated, even concomitantly with the requisite background state C, without the conscious state arising? (I should note here that Dave argues for this not from his thesis that any physical state can be separated from any conscious state because consciousness is not reducible to physicality, but rather from the hypothetical situation of 'downstream lesions', lesions that cut off an NCC from other, later processing that might result in a loss of consciousness, even though the NCC was intact and in state N, Sarah has something to say against this possibility, but even if her argument is correct, Dave could always rely on his earlier claim of separability.) Anyway, the point I want to make is that if we take these two possibilities seriously, from a practical standpoint we might never arrive at the 100% correlation that necessity, sufficiency or even minimal sufficiency demands. Mightn't we be better off (epistemically) if we were to take a reliabilist stance, and judge putative NCCs by their correlation percentages, positive and negative, with conscious states? So, for example, NCC1 might have a 85% positive correlation with Conscious State 1 (CS1), meaning that 85% of the time that NCC1 is in N, CS1 occurs, and a 90% negative correlation, meaning that 90% of the time that NCC1 is not in state N, CS1 does not occur. Of course, these ratios can come apart because of the possibilities of multiple realizability and full physical instantiation without phenomenal (conscious) instantiation. Then we could compare different candidate NCCs by seeing how their numbers stack up, and we could figure out what sort of shortcomings an NCC might have given its different performances on the two scales, e.g. an NCC with a high positive but low negative correlation might be one of a number of individually sufficient NCCs for that CS. Also, we could use these scores to settle questions about where to draw the boundaries of an NCC, by referring to the changes in the numbers that occur when we include (or exclude) certain substructures of an NCC. For these reasons, I think it might behoove us to focus on this probabilistic correlation as the defining one of NCCs, and not the all-or-nothing one we have been using so far.

My second worry is over Dave's suggestion that artificial brain stimulation might be used to exclude non-core elements of an NCC, in the following way: If we can artificially stimulate a neural network 'downstream' from a putative NCC, and we get the same CS as we did when the NCC was supplying the inputs, then we have a case for eliminating the earlier NCC, since it no longer seems necessary for the CS. What worries me here is that it seems that what we are doing when we 'artificially stimulate' the downstream system is taking on the functional role of the earlier system. But if this is the case, why is it impossible to take on the functional role of part of an NCC? And if it is possible, how will we differentiate between 'merely' background or enabling states and actual

(parts of) NCC states? It seems to me that unless there is some a priori reason that a machine cannot assume the functional role of part of an NCC, then we have to come up with a way to tell them apart, otherwise we risk arbitrarily excluding neuronal groups that are possibly part of an NCC.

Which brings me to my third point. One might argue that one way to tell a background or non-core neural state from an NCC one is via temporal indexing, using the onset of consciousness as a benchmark. So, we might say, anything that occurs prior to the onset of consciousness is a possible background system, but not part of an NCC, and anything that occurs afterward is at least possibly part of one. But this is insufficient in two ways: Firstly, even if it were a viable method of parsing, it would only answer questions about upstream neural candidates, and second, it doesn't even seem a viable method, to me, because it tacitly assumes a symmetry between conscious temporality and physical temporality. If we are going to use the onset of consciousness as our temporal benchmark, and exclude neural candidates for NCChood on the basis of their preceeding this benchmark, then we have to assume that the onset of consciousness is co-temporaneous with the NCC processing that is responsible for it, and this is not necessarily the case. It might be that the onset of consciousness lags behind NCC processing, and as a result the subject enters the CS after the crucial NCC processing has occurred. We can imagine further that some other processing, downstream of the NCC, occurs co-temporaneously with the onset of consciousness. In such a scenario (which seems completely possible to me) if we were to employ the benchmarking method outlined above, we might well associate the CS with the later, non-NCC processing, because the actual NCC processing occurs before the onset of consciousness, and is therefore 'upstream'. As to why consciousness might lag behind the NCC processing that underlies it, who knows, but the relationship between the physical and the phenomenal is certainly mysterious enough for this to be possible.

(I apologise to Dave for the lateness and to everybody for the long-windedness and pedantry)
Al.

From sawright@U.Arizona.EDU Sun Sep 26 19:31:38 1999
Return-Path: <sawright@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id TAA17287
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 19:31:37 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGFRJXCXLT8B8VIC9@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 26 Sep 1999 19:31:47 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTTP id <01JGFRJBBX28B8VZMG@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 26 Sep 1999 19:31:44 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id TAA17282 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 19:31:18 -0700
Received: from localhost (sawright@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTTP id TAA20849 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 19:31:29 -0700 (MST)
Date: Sun, 26 Sep 1999 19:31:29 -0700 (MST)
From: Sarah A Wright <sawright@U.Arizona.EDU>
Subject: Downstream Lesions
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909261927060.19732-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

I wanted to dwell a bit on the worries about lesions in "What is the Neural Correlate of Consciousness?" and from last class. It seems that lesion studies should be scrutinized with care both because they can lead us to exclude some of the NCC from our NCC, and because they can lead us to include some non-NCC in our NCC. (They seem to make the NCC both too wide and too narrow.)

In the first case, we have the worries about plasticity of the brain. If, after a lesion, the brain is capable of rerouting its processing in such a way that the NCC actually changes location, then the information from the study of such lesions might give us a sufficient NCC, but not the NORMAL sufficient NCC. Relying only on lesion studies might cause us to reject as the NCC an area of the brain that we normally use as the NCC, but which in marginal cases we can work around.

In the second case, we have the putative NCC in a petri dish, or other less extreme versions of this problem. In these cases it seems that the putative NCC would not be conscious. However we don't want to hold this against the NCC we have isolated, since there are certain background conditions we want to hold fixed in our search for the NCC. These background conditions might include getting appropriate (not processed) sensory data, or, if there is an on switch for consciousness, having that switch on. Requiring the NCC to operate despite all lesions (down to amputation) seems to ignore the fact that we are looking for the MINIMAL sufficient NCC.

Now the impetus for each of the worries above seem to be directed at lesions that occur upstream of (or in) the NCC. Sensory input should occur far upstream of the NCC, and, whatever an "on switch" might turn out to be it should be activated either before or at the time of consciousness. Likewise, it seems that lesions before or in the NCC are the ones we could use plasticity to get around.

These concerns have brought us to the the question of where to draw the upstream line on the NCC, or how far back we need to go. There is another related question of how far forward we should draw the downstream line of the NCC.

In particular I was bothered by this claim in "What is the Neural Correlate of Consciousness?"

>Less radically, one can imagine placing lesions immediately downstream
>from a candidate NCC N, so that N's effects on the rest of the brain are
>significantly reduced. In such a case, it is probable that N can be
>active without the usual behavioral effects associated with
>consciousness, and *quite plausibly without consciousness itself*.

I am willing to admit that the downstream limit of the NCC need not be as far down as behavior; we certainly don't want to say that someone who is paralyzed cannot have a full NCC simply because they cannot report it. But it seems to me that a lesion downstream of the NCC should not interfere with consciousness itself.

In the petri dish case we discount certain background conditions because while we can't have consciousness without them, they also don't seem relevant to conscious experience. Sensory input and brain "on"ness are plausible examples. But what could the background conditions here be? These would be conditions that occur after consciousness, but also enable it. I cannot think of any plausible examples of this sort of background condition. If an event downstream from our NCC is needed for consciousness, rather than discount this event, we should update our NCC to include the substrate of that later event. (Note that the background

conditions needed here aren't those (epistemological ones) that make the consciousness of others available to us, but rather those (ontological ones) that allow conscious experience.)

If you all can think of any examples of a downstream background condition that should be discounted, please let me know.

From lan@U.Arizona.EDU Sun Sep 26 22:20:56 1999
Return-Path: <lan@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA17539
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 26 Sep 1999 22:20:55 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
id <01JGFXG9R51CB8UKZ0@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 26 Sep 1999 22:21:05 MST
Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDf V5.1-12 #24137)
with ESMTP id <01JGFXG8P4JKB8VZUA@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Sun, 26 Sep 1999 22:21:04 -0700 (MST)
Received: from localhost (lan@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id WAA26118; Sun,
26 Sep 1999 22:21:03 -0700 (MST)
Date: Sun, 26 Sep 1999 22:21:03 -0700 (MST)
From: Lonnie A Nelson <lan@U.Arizona.EDU>
Subject: Considering conditions C
In-reply-to: <199909230807.BAA09334@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909252056340.9865-1000000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Given that the necessity and sufficiency issue has lead the discussion to speculate about possible versions of "conditions C" which would be the "normal conditions under whcih we would expect a NCC to operate". The question has become "what are normal conditions?" As Dave mentioned in class one could start as far away from the brain as the type of stimulus that one is exposed to and only allow "environmentally valid" stimuli. This has coherence under the "natural functioning" band of thought, but it won't really allow us to play with the conscious/subliminal distinction that would need to be experimentally allowed in order to have any data whatsoever to procede with.

While I believe that we should leave the brain and perceptual systems in as pristine a state as possible so that we can observe the system's "normal functioning", I would question what is referred to as "environmentally valid" regarding the allowed for stimuli set. While binocular rivalry may be a supremely unlikely situation to find oneself in, we have in no way said that the stimuli needed to be "evolutionarily valid" as the only monkeys finding themselves needing to discern binocular rivalrous stimuli are in a lab with electrodes in their brains, hence not a "naturally occurring" condition. However, if we take Environmentally valid, and make it into "natural reception system". thus cutting the world off, for the time being at the level of the organism, we have a reasonably wide array of possible manipulations that we are able to subject the system to and that we may monitor its activity in response to.

For these "natural modes of signal reception" we have (as should be obvious) our five senses. The most easily controlled of the phenomenal consciousness set. The case being that psychophysics has given us a pretty good idea of general thresholds for the various senses with regard to their respective stimuli; we should be able to monitor the difference

between JND (Just noticeable difference) below and above the threshold for conscious experience. The difference in activity of the system between these two inputs should give us a reasonable approximation of the brain subsystems involved in "phenomenal experience" of a selected modality. Given that the actual energy difference in the light input for the two stimulus strengths can be known, a subtraction from the output values (in units expressed by activity) could be used to determine the generative potential of conscious experience, and the localization of this activity can be carried out via PET imaging, or fMRI (as long as you aren't using audition as your mode of exploration). And while these imaging techniques certainly have their drawbacks in terms of fine versus rough grained localization, we are given the added benefit of not having to disturb the system or "sacrifice" it post experiment to make sure that we were measuring the right (fine grained) place.

While this methodology will surely only give us a rough picture of a NCC it will also be an accurate one in the "natural state of the system", which may be of some extra value, given that we would not have to deal directly with the plasticity problem.

Just some thoughts,

--Lonnie

It is a common fate of all knowledge to begin as heresy and end as orthodoxy.

-Thomas Huxley

Lonnie A Nelson
Department of Psychology
Human Energy Systems Laboratory
University of Arizona
lan@u.arizona.edu

From lachter@u.arizona.edu Mon Sep 27 10:45:27 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id KAA18548
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 10:45:27 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGGNGD03Y8B8R43M@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 27 Sep 1999 10:45:37 MST
Received: from trifid.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JGGNGBR23KB8VV3N@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Mon, 27 Sep 1999 10:45:36 -0700 (MST)
Received: from [128.196.99.98] ([128.196.99.98])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id KAA26716; Mon,
27 Sep 1999 10:45:34 -0700
Date: Mon, 27 Sep 1999 10:45:45 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: Considering conditions C
In-reply-to: <Pine.HPX.4.10.9909252056340.9865-100000@pavo.U.Arizona.EDU>
X-Sender: lachter@pop.u.arizona.edu
To: Lonnie A Nelson <lan@u.arizona.edu>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a02b41558b08e68@[128.196.99.98]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
References: <199909230807.BAA09334@paradox.soc-sci.arizona.edu>
Status: RO

Lonnie proposes that we look for NCCs by doing imaging studies where we

subtract visual distinctions which are just below one just noticeable distance (JND) from those that are just above it. While this seems plausible in principle there are a couple things to keep in mind. First, notions like "JND" and "threshold" are really statistical. They are defined as the point where an observer can get some percentage of trials correct (usually 75%). Thus, just below a JND, the subject will usually be getting 70% correct and just above they will be getting 80%. The whole system is graded, and the parts necessary for consciousness are not likely to be much different.

Second, you are likely to get bigger differences as you move downstream independent of the location of any NCC. This is because all that downstream stuff that is hanging out thinking "Is that something? Where is it? Oh, I think that was it? Maybe not..." is going to act differently when it actually is getting information from when it is just hanging out.

Third, it is not clear that objective and subjective thresholds are the same or should be measured the same way (you are likely to get a lot of flack on this no matter how you do it).

In short, while this is an appealing idea, I don't think it could actually be made to work.

Joel

From lan@U.Arizona.EDU Mon Sep 27 14:00:14 1999
Return-Path: <lan@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id OAA18930
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 14:00:13 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGGU9S73LCB8VXUE@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 27 Sep 1999 14:00:23 MST
Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JGGU9QXJNKB8LOKA@Telcom.Arizona.EDU> for
chalmers@Ariziona.EDU; Mon, 27 Sep 1999 14:00:20 -0700 (MST)
Received: from localhost (lan@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id OAA27715; Mon,
27 Sep 1999 14:00:18 -0700 (MST)
Date: Mon, 27 Sep 1999 14:00:17 -0700 (MST)
From: Lonnie A Nelson <lan@U.Arizona.EDU>
Subject: Re: Considering conditions C
In-reply-to: <v04020a02b41558b08e68@[128.196.99.98]>
To: Joel Lachter <lachter@U.Arizona.EDU>
Cc: David Chalmers <chalmers@Ariziona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909271351490.20674-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

On Mon, 27 Sep 1999, Joel Lachter wrote:

I was aware that the threshold was a statistical construct, however, it should not take all that long to test for each subjects individual threshold (perhaps 30 trials, to assume a random distribution). and you would still have a difference even if there were only 30% that the subject was "not seeing" you could compare imaging on that 30% to the trials in which they "saw" the stimuli

> Second, you are likely to get bigger differences as you move downstream

> independent of the location of any NCC. This is because all that downstream
> stuff that is hanging out thinking "Is that something? Where is it? Oh, I
> think that was it? Maybe not...", is going to act differently when it
> actually is getting information from when it is just hanging out.

Wonderful, these are the differences that we will be looking for. What we
want to subtract is the "just hanging out" from the "getting information",
the larger the differences, the happier we should be. One would expect
that the largest difference should be able to be found in the actual NCC
itself, as it "got new information" and is no longer "just hanging out".

> Third, it is not clear that objective and subjective thresholds are the
> same or should be measured the same way (you are likely to get a lot of
> flack on this no matter how you do it).

This can be done in some way similar to the first reply I made (see
above) on individual thresholds. And when you are trying to do something
in a novel way, you will always catch flack, the object should be to get
the most valid form of information that you can for each subject that you
test. and While I realize that this sounds easier than it would be in
actuality, this is not so big a problem that the whole process could not
be made to work, imho.

--Lonnie

It is a common fate of all knowledge to begin as heresy and end as
orthodoxy.

-Thomas Huxley

Lonnie A Nelson
Department of Psychology
Human Energy Systems Laboratory
University of Arizona
lan@u.arizona.edu

From lachter@u.arizona.edu Mon Sep 27 16:06:31 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA19051
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 16:06:31 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGGYOESWSWA63QBN@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 27 Sep 1999 16:06:41 MST
Received: from trifid.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JGGYODBC8WB8W6LE@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Mon, 27 Sep 1999 16:06:39 -0700 (MST)
Received: from [128.196.99.98] ([128.196.99.98])
by trifid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id QAA17456; Mon,
27 Sep 1999 16:06:37 -0700
Date: Mon, 27 Sep 1999 16:06:38 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: Considering conditions C
In-reply-to: <Pine.HPX.4.10.9909271351490.20674-100000@orion.U.Arizona.EDU>
X-Sender: lachter@pop.u.arizona.edu
To: Lonnie A Nelson <lan@u.arizona.edu>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a04b4159b0cdb0b@[128.196.99.98]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
References: <v04020a02b41558b08e68@[128.196.99.98]>
Status: RO

At 2:00 PM -0700 9/27/99, Lonnie A Nelson wrote:

>I was aware that the threshold was a statistical construct, however, it
>should not take all that long to test for each subjects individual
>threshold (perhaps 30 trials, to assume a random distribution). and you
>would still have a difference even if there were only 30% that the subject
>was "not seeing" you could compare imaging on that 30% to the trials in
>which they "saw" the stimuli

I don't know how much it is worth belaboring this point but...

The problem is that you can't pick out the trials on which subjects saw the stimulus from the trials on which they did not. When you are performing such a task you feel like you are guessing on every trial, and the computer is often telling you you are wrong on those few trials you feel somewhat confident about. The correct trials are contaminated by ones where you were just guessing. The wrong trials are contaminated by ones where you thought you saw something but answered the other way because the last three times you responded that you saw it when you were not completely sure you got it wrong. It is not the case that you either see these stimuli or you don't. There is a gray area. When you have stimuli near threshold you are in that gray area.

>One would expect
>that the largest difference should be able to be found in the actual NCC
>itself, as it "got new information" and is no longer "just hanging out".

I would not expect it to turn out that way. Some might say that that almost guarantees that it would turn out that way:-) However, it seems to me that any plausible NCC is going to be in some gray area, reflecting how well the observer can actually see the stimulus. Areas farther downstream, however are going to be acting differently because they are not providing a graded output. Downstream areas will act in qualitatively different ways depending on how much information they have (eg., little information = frustration, thoughts of breaking computer; midlevel = interest, pride when you get it right, anger when you get it wrong; higher level = boredom, can do in sleep, thoughts of lunch). In all probability the NCC does not think "Did I see something?" It just represents the available information. Stuff farther downstream thinks about this information and thinks and acts in entirely different ways depending on the amount of information available.

Joel

From lan@U.Arizona.EDU Mon Sep 27 18:56:55 1999

Return-Path: <lan@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id SAA19565

for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 18:56:54 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JGH4MOP0W0B8WJRL@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Mon, 27 Sep 1999 18:57:05 MST

Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

with ESMTTP id <01JGH4MN7NZKB8WB2T@Telcom.Arizona.EDU> for

chalmers@Ariziona.EDU; Mon, 27 Sep 1999 18:57:04 -0700 (MST)

Received: from localhost (lan@localhost)

by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTTP id SAA13335; Mon,

27 Sep 1999 18:57:02 -0700 (MST)

Date: Mon, 27 Sep 1999 18:57:02 -0700 (MST)

From: Lonnie A Nelson <lan@U.Arizona.EDU>

Subject: Re: Considering conditions C

In-reply-to: <v04020a04b4159b0cdb0b@[128.196.99.98]>

To: Joel Lachter <lachter@U.Arizona.EDU>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909271842030.8354-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

On Mon, 27 Sep 1999, Joel Lachter wrote:

> I don't know how much it is worth belaboring this point but...

I am relatively sure that this will not have any far reaching impact on the state of the science, but...

> The problem is that you can't pick out the trials on which subjects saw the
> stimulus from the trials on which they did not. When you are performing
> such a task you feel like you are guessing on every trial, and the computer
> is often telling you you are wrong on those few trials you feel somewhat
> confident about. The correct trials are contaminated by ones where you were
> just guessing. The wrong trials are contaminated by ones where you thought
> you saw something but answered the other way because the last three times
> you responded that you saw it when you were not completely sure you got it
> wrong. It is not the case that you either see these stimuli or you don't.
> There is a gray area. When you have stimuli near threshold you are in that
> gray area.

There should, then be a place on either side of that gray area, one of which is above, and one of which is below or within that gray area.

> >One would expect
> >that the largest difference should be able to be found in the actual NCC
> >itself, as it "got new information" and is no longer "just hanging out".
>
> I would not expect it to turn out that way. Some might say that that almost
> guarantees that it would turn out that way:-) However, it seems to me that
> any plausible NCC is going to be in some gray area, reflecting how well the
> observer can actually see the stimulus.

Which is why the Above condition would have to be "above" and the other would have to be within the grey or below it.

> Areas farther downstream, however
> are going to be acting differently because they are not providing a graded
> output. Downstream areas will act in qualitatively different ways depending
> on how much information they have (eg., little information = frustration,
> thoughts of breaking computer; midlevel = interest, pride when you get it
> right, anger when you get it wrong; higher level = boredom, can do in
> sleep, thoughts of lunch).

THIS could undoubtedly be made at least a little bit interesting, the vast majority of cognitive research is really boring for the subject, but people participate and effects are found. If there is sufficient difference between trials that can be seen with high accuracy (~100% +/- 3%) vs. slightly below threshold (~chance) the task would not necessarily be that difficult or frustrating, (though boredom will definitely be a factor, but hopefully thoughts of lunch would be randomly distributed across your sample:-) and the imaging should still tell us the differences in the reactivity of the brain. THOUGH we would likely get a picture of the processing going on afterward, this is arguably part of the reaction of the NCC.

> In all probability the NCC does not think "Did I

> see something?" It just represents the available information. Stuff farther
> downstream thinks about this information and thinks and acts in entirely
> different ways depending on the amount of information available.

>

Maybe we are drawing the line in different places for an NCC, I have no
desire to cut off the "higher order processes" of consciousness, only to
separate out the difference between conscious percept and *not*.

Lonnie

It is a common fate of all knowledge to begin as heresy and end as
orthodoxy.

-Thomas Huxley

Lonnie A Nelson
Department of Psychology
Human Energy Systems Laboratory
University of Arizona
lan@u.arizona.edu

From lachter@u.arizona.edu Mon Sep 27 21:32:03 1999
Return-Path: <lachter@u.arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA19710
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 27 Sep 1999 21:32:03 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGHA21BZ80B8W05N@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Mon, 27 Sep 1999 21:32:14 MST
Received: from trivid.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
with ESMTP id <01JGHA2077XCB8VX3H@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Mon, 27 Sep 1999 21:32:13 -0700 (MST)
Received: from [10.0.2.15] (tec3.Psych.arizona.edu [128.196.98.11])
by trivid.u.arizona.edu (8.8.8/8.8.8) with ESMTP id VAA15192; Mon,
27 Sep 1999 21:32:09 -0700
Date: Mon, 27 Sep 1999 21:31:48 -0700
From: Joel Lachter <lachter@u.arizona.edu>
Subject: Re: Considering conditions C
In-reply-to: <Pine.HPX.4.10.9909271842030.8354-100000@orion.U.Arizona.EDU>
X-Sender: lachter@pop.u.arizona.edu
To: Lonnie A Nelson <lan@u.arizona.edu>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <v04020a01b415e689c1cc@[128.196.99.98]>
MIME-version: 1.0
Content-type: text/plain; charset=us-ascii
References: <v04020a04b4159b0cdb0b@[128.196.99.98]>
Status: RO

At 6:57 PM -0700 9/27/99, Lonnie A Nelson wrote:

>There should, then be a place on either side of that gray area, one of
>which is above, and one of which is below or within that gray area.

Yeah. But now you are not talking about "just above one JND" and "just
below one JND". I take it the force of your original proposal was that you
could have two stimuli which looked very similar to the retina, but which
look quite different consciously and thus presumably quite different at the
NCC. Once you move things out of the gray area you are talking about things
that look quite different on the retina.

>Maybe we are drawing the line in different places for an NCC, I have no
>desire to cut off the "higher order processes" of consciousness, only to
>separate out the difference between conscious percept and *not*.

It seems to me (tell me where I am wrong) that you are proposing running an experiment with two conditions. In condition A the subject can see the stimuli. In condition B the subject cannot see the stimuli. We then do a subtraction to see where "seeing" is occurring. If I have reconstructed your proposal correctly I see two problems with it. One is that you cannot make condition A and condition B comparable on the low level visual stuff (and still keep your clean distinction between being able to see the stimuli and not being able to see the stimuli). As a result, low level things will light up in your subtraction. Similarly, you cannot keep condition A and condition B comparable with respect to high level stuff. Doing a task when you can see the stimuli from doing it when you cannot see the stimuli in more ways than simply being able to see the stimuli. These conditions also differ in cognitive and emotional respects which are likely to cascade throughout the system. So the problem is, before running the experiment, I don't have any strong intuition telling me that the neural correlate of seeing these stimuli is going to light up more strongly than, say, the neural correlate of attending to the banging of the magnet, or the neural correlate of fidgeting. Thus I don't see how you can tell after running the experiment whether you have found the neural correlate of seeing vs a myriad of other possibilities.

Joel

From chalmers@paradox.soc-sci.arizona.edu Tue Sep 28 23:31:14 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA22099
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 28 Sep 1999 23:31:14 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGISI6JUHC8WY8J@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 28 Sep 1999 23:31:27 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGISI5BJ00B8W02F@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Tue, 28 Sep 1999 23:31:25 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id XAA22086 for scicon; Tue, 28 Sep 1999 23:30:40 -0700
Date: Tue, 28 Sep 1999 23:30:40 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: NCC postings
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199909290630.XAA22086@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R

Hi all, good to see the spate of interesting NCC postings. Here are a few brief comments:

Overall comment: we have to distinguish the three questions: (i) what is it to be an NCC, (ii) is there likely to be an NCC, and (iii) what will an NCC tell us about consciousness?

In my own paper, I was mostly trying to give a plausible answer to (i) based on the way the notion is used in the field. I also said a few words about (ii), suggesting that it wasn't unreasonable to expect that there may turn out to be NCCs so defined. I said very little about (iii).

As I said in the paper, I think it is strategically useful for people in the field to define NCC in terms of correlation alone, without any

additional claims such as, the NCC is the basis of consciousness, the NCC generates consciousness, the NCC explains consciousness, the NCC is the key to a theory of consciousness. That way, the search for NCCs is a common, relatively "objective" project that almost everyone can participate in without too much ideological baggage. Of course at the end of the day, that does leave open the question of what, if anything, the NCC will tell us about consciousness.

LIS: I take it that Lis is mostly concerned with (iii). The worry is that physiology alone doesn't tell us much; it will only tell us something in the context of a theory. I think there's a lot to be said for that; on the other hand, it is quite possible that isolating an NCC will itself help in the development of a theory. To take Li's example of binocularly rivalry, learning that IT is an NCC doesn't tell us anything much interesting about binocularly rivalry per se. But once we know this, we can look at how and when IT is activated, and how it related to other brain areas, and perhaps reach some very interesting conclusions about the place of conscious vision in binocular rivalry.

I do agree with a lot of what Lis says about Crick & Koch and Logothetis's approach being based on assumptions (re temporality, re connections to prefrontal cortex) that may not be justified. The trouble is that everyone needs to make some assumptions to get these things off the ground. Some assumptions are more reasonable than others; when they seem weak, one had better hope that the assumption itself will ultimately be testable. Of course I think that at the end of the day there are some pre-experimental background assumptions that everyone needs to rely on, but these had better be pretty straightforward. (E.g., the connection between report and consciousness.)

Lis points that consciousness might somehow be grounded in neural areas independent of content, and that content just "feeds into" consciousness, is interesting. I guess there still has to be a distinction between those contents that make it into consciousness and those that don't; it's not as if merely representing a content anywhere in the brain gets it into consciousness. So one will still have a relevant distinction between content-related neural systems here. Then the further question is whether the content has to "go somewhere" (e.g. to Lis's base system) to make it into consciousness. If that's the case, then content representation in the base system will be relevant after all. If that's not the case, then (as Simon points out) it's hard to see just why some contents make it into consciousness and not others.

NICK: Nick says that a potential NCC N may not be sufficient on its own for consciousness, since one needs the right kind of activity in N. I think that's right, and strictly speaking an NCC should be characterized also in terms of the sort of activity. But the sort of general point I was making will still hold -- one can imagine that exactly that sort of activity could be happening inside N in a petri dish, and one presumably wouldn't get consciousness. So even this activity in N isn't sufficient for consciousness, except against the background of conditions C. The relations to other brain areas that Nick talks about might in effect be part of conditions C (normal functioning brain).

JURAJ: Juraj makes the excellent point that the sort of questions I raise (re lesions etc.) about NCCs also come up for neural correlates of other sorts of mental states and conditions, such as prosopagnosia

(and presumably learning, memory, language, etc, etc, etc). In fact they arguably come up throughout biology, engineering, etc, when we are concerned with finding the "area for X", the "mechanism of X", and so on. I don't know if there is much literature on the general problem, but there ought to be. And maybe some conditions, such as prosopagnosia, will raise special problems (for the reasons Juraj suggests), so one will only be able to identify correlates in a coarse-grained way. Maybe the thing to say would be that in these cases, there is just no fact of the matter about what the "real correlate" is -- there's just a complex causal chain.

LOGAN: Logan raises the interesting possibility that early visual areas may produce "low-level" qualia while later areas produce "high-level qualia". A normal visual experience has elements of each and so has correlates in both places. Cases such as intermediate blindsight may have the info in the later areas but not the early areas, yielding high-level qualia without the low-level qualia, which would be understandably strange!

I guess, if the high-level qualia are "cognitive qualia", we have interpretation (3) (I think) of intermediate blindsight from today, ie. conscious visual judgment but no sensory phenomenology. On the other hand, if the high-level qualia are more like high-level sensory phenomenology, something like the grouping of the pixels, or the "seeing" of the light patterns as a dalmatian, then we'd have something closer to (i) or (ii). Understandably, this could be very strange.

PETER: Peter asks, how can IT and not V1 be the NCC, but IT gets input from V1? I guess the answer is that the NCC could come at an intermediate stage (because e.g. it takes processing for info to become conscious), in which case, it's only to be expected that it will get input from areas that aren't themselves part of the NCC.

Of course there may well be multiple NCCs, as Peter says. I think this actually answers Peter's other question, of why we only require sufficiency and not necessity of an NCC. If there are multiple NCCs, any one won't be necessary. But sufficiency alone will mean that it's "important" for consciousness in some sense -- activate this system, and we get consciousness! Of course, one still has to understand how and why the NCC correlates with consciousness, but sufficiency might be taken as a sign that we're at least looking in the right ballpark.

JOEL P.: Joel suggests that split brain studies give reasons to believe in multiple NCCs. These people might be interpreted as having two streams of consciousness, and two NCCs. Normal people like us presumably have both neural systems, so why don't we have two consciousnesses two? I guess one could get the beginnings of an answer by saying that the difference in normal functioning between ourselves and split brain cases is such that there's no reason to believe that the NCC for one will also be the NCC for the other -- it's just the sort of case where we don't want to draw conclusions about our own NCC from lesion studies!

Personally I'm not at all sure that split brain cases involve two streams of consciousness. But let's say they do, and that they have two corresponding NCC's. My own guess would be that nevertheless, we don't have two corresponding NCCs, but rather one corresponding NCC that goes across both hemispheres. Most of the time, the info in the two hemispheres is in synch, so we get a coherent consciousness; if the info were to come radically apart, we might have a weird consciousness, but still within a single consciousness. Basically,

the reason would be that streams of consciousness might go with something like loci of integration and integrability. In the split brain cases, the info in the two NCCs is not integrable with each other, so we get two streams. In our case, info is integrable (e.g. across the corpus callosum), so we get one stream.

SIMON: Simon makes interesting points both re what Lis says about content (as above) and about lesions. The weaker interpretation of lesion studies will basically be that if lesioning area N affects visual consciousness, then area N will be involved in the causal chain underlying visual consciousness. That seems like a reasonable conclusion. Though of course for all we've said it could be a fairly distant and indirect part of the causal chain, in the way that the heart is!

SARAH: Sarah worries about how downstream lesions could change the location of the NCC. I admit my remarks here were a bit brief and cryptic. But here's one way things might work. Let's say that the NCC will be the locus of direct availability for global control, as I suggest, or the mechanism of the global workspace, as Baars suggests. Then it may be that if one lesions downstream from the NCC, this system will no longer be the locus of direct availability for global control (it won't play the control role any more), and it won't be the global workspace (as it won't transmit the info any more). Maybe other areas will take over and fill the void instead, they'll send forward the info that controls behavior, etc. If so, the original system will no longer be the NCC, and other systems will become the NCC.

Something like this will happen if we think that what makes the NCC an NCC is in part its functional role, or at least its downstream functional role (whatever that role is). Downstream lesions may kill off that role, and so may kill off the fact that the system is an NCC. Of course, if one doesn't think that downstream functional role is relevant to making the NCC an NCC, this won't follow.

ALLEN: Allen suggests, as does Logan, that maybe weak (less than 100%) correlation might be enough to qualify a system as an NCC. My main worry here is that this way, too many brain systems might qualify as NCCs. E.g., the retina state, V1 state, etc, might all fairly reliably correlate with consciousness, and come apart only in special circumstances (e.g. binocular rivalry). It seems that in such cases we want to say there's something special NCC-wise about the system that still correlates as opposed to dissociating. So I think maybe there's a case for requiring full correlation. If a system sometimes dissociates from consciousness, then we should be looking either elsewhere or to a broader system for an NCC.

Allen also raises the very interesting point that artificial brain stimulation might make another system M take on the functional role definitive of the NCC, where system N had played it before. In this case, we'll dissociate N from consciousness, suggesting that N isn't really the NCC; but maybe we want to say is that in this case we've actually moved the NCC from N to M, because of the stimulation. Maybe this is another reason for being suspicious of artificial brain stimulation as a criterion, just as we are suspicious of lesions. At least, for being suspicious of too much stimulation. If it's just a little stimulation, making M active in a normal way when it hadn't been a moment ago, it's hard to see how this could make M play a functional role that it couldn't normally play. What would be required for Allen's scenario is stimulation that makes M active in an

abnormal way, thus making it play an abnormal functional role. Maybe we should have some sort of restriction to stimulation of "normal" patterns of activity, if that makes any sense.

Allen also raises the worry that if consciousness lags behind NCC activity, the temporal matching methodology may go wrong. That's a good point, and in synch with what Lis says about the temporal methodology being based on assumptions. One question is whether the assumption could itself be tested, e.g. by finding correlations based wholly on non-temporal aspects, and then seeing how the temporal aspects line up. If they don't, then we'll have to be careful! In advance of doing that, presumably any conclusions based only on temporal matching will have to be fairly tentative.

LONNIE AND JOEL L.: Lonnie suggests that one route to finding the NCC might involve imaging the brain across closely related conditions of conscious and unconscious perception, separated by a JND or some such. Joel points out that around the "threshold" one is really almost arbitrarily guessing that one saw it or not, so it may not correspond to a deep conscious/unconscious distinction.

One potential way to avoid this problem is to give up on the JND idea, and to image relevantly similar episodes of clearly conscious vs. clearly unconscious perception. E.g., that is what Weiskrantz, Zeki, and others have done with blindsight patients such as G.Y. Image him in "blind" seeing, image him in standard seeing, and see what brain areas are different. The result is that a lot of different brain areas show up as different, as Joel predicted. Presumably one reason is that a difference in consciousness will also be associated with a lot of differences "downstream" from consciousness, e.g. in language, planning, thinking, etc, and related areas. Zeki does claim to find some fairly consistent relevance of V5, though.

ADAM: Adam worries that the whole NCC approach is premature, since we need to get a good phenomenological method off the ground first. Not just to gather the first-person data, but also to gather the third-person data from neuroscience, etc. My worry here is that the same presumably goes for third-person data in any science at all, e.g. physics, chemistry, biology. But if Newton had waited for phenomenology to be properly developed, we would have been waiting a long time! My own view is that a developed phenomenology is vital for gathering first-person data, and may end up helping us gather third-person data, but that at least the third-person part of the story ought to be able to get off the ground without it. It's not clear why the NCC search is any worse off here than any other area of neuroscience, or of science in general.

--Dave.

From rachael@U.Arizona.EDU Thu Sep 30 23:29:25 1999

Return-Path: <rachael@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA25995

for <chalmers@paradox.soc-sci.arizona.edu>; Thu, 30 Sep 1999 23:29:25 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JGLL00I0DCB8XB9X@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Thu, 30 Sep 1999 23:29:40 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU

(PMDF V5.1-12 #24137) with ESMTP id <01JGLL0MKWPCB8XE8K@Telcom.Arizona.EDU>

for chalmers@Arizona.EDU; Thu, 30 Sep 1999 23:29:38 -0700 (MST)

Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA25990 for

<scicon@paradox.soc-sci.arizona.edu>; Thu, 30 Sep 1999 23:29:12 -0700
Received: from localhost (rachaelp@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id XAA26869; Thu,
30 Sep 1999 23:29:29 -0700 (MST)
Date: Thu, 30 Sep 1999 23:29:29 -0700 (MST)
From: Rachael J Parkinson <rachaelp@U.Arizona.EDU>
Subject: more on NCCs
In-reply-to: <19990930231157.11402.qmail@hotmail.com>
To: chalmers@U.Arizona.EDU
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909302312240.22861-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Just one more thing on NCCs...

In "What is the Neural Correlate of Consciousness?" Chalmers discusses what sorts of conscious states would be relevant to determining the neural correlate of consciousness. He explores three classes of states that we should consider: that of being conscious (or unconscious), background states of consciousness (including sleep, hypnosis, dreaming, etc.) and contents of consciousness. It is the latter, contents of consciousness, that Chalmers argues may be the most interesting states of consciousness. It is these that I am concerned with in this post.

It seems to me that, with the right methodology and experimentation, we may come to discover neural correlates of vision, for example. Certainly the work of Logothetis and others points to something positive in that direction. Suppose then, that we were to discover a correlate between particular 'horizontal' neurons and first-person 'horizontal' experience. Though I agree that this would be a significant step in developing a Chalmersian science of consciousness, I would be reluctant to identify such a finding as the 'neural correlate of consciousness.' This is because it is inadequate to the data of our phenomenal experience which rarely, if ever, consists of the single experience of a horizontal line. Our phenomenal experience is rich and diverse and it somehow comes together in our unified subjective experience. It seems to me that a neural correlate of consciousness then, must at least point to what unifies our phenomenal experience. If it does not, then it seems that what we are looking at is merely the 'neural correlate of vision' or the 'neural correlate of audition.'

Note that I am not advocating any sort of a priori claims about a Cartesian theater. It may well be that the unity of consciousness can not be attributed to any particular set of neurons or brain processes. But it seems that what we should be looking for in our search for the neural correlate of consciousness is something that, though not offering a reductive explanation, can account for, or point to, that thing which correlates with our *unified* phenomenal, conscious experience.

What do others think on this matter?

Best,
Rachael

From landsurveyor@hotmail.com Thu Sep 30 16:12:13 1999
Return-Path: <landsurveyor@hotmail.com>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA25678
for <chalmers@paradox.soc-sci.arizona.edu>; Thu, 30 Sep 1999 16:12:13 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGL5QMZ80WB8WSF4@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Thu, 30 Sep 1999 16:12:28 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGL5QLPKCGB8X9HX@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Thu, 30 Sep 1999 16:12:26 -0700 (MST)
Received: from hotmail.com (f201.hotmail.com [207.82.251.92])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with SMTP id QAA25673 for
<scicon@paradox.soc-sci.arizona.edu>; Thu, 30 Sep 1999 16:11:50 -0700
Received: (qmail 11403 invoked by uid 0); Thu, 30 Sep 1999 23:11:57 +0000
Received: from 128.196.51.170 by www.hotmail.com with HTTP; Thu,
30 Sep 1999 16:11:57 -0700 (PDT)
Date: Thu, 30 Sep 1999 16:11:57 -0700 (PDT)
From: Matt Herbert <landsurveyor@hotmail.com>
Subject: blindsight and access
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <19990930231157.11402.qmail@hotmail.com>
MIME-version: 1.0
Content-type: text/plain; format=flowed; CHARSET=US-ASCII
X-Originating-IP: [128.196.51.170]
Status: RO

Block says that a blindsight patient is both phenomenal- and access-unconscious. This runs contrary to what some theorists think. The striking fact about blindsight is supposed to be that it is a case of access without phenomenal experience. The patient makes an inference about x without experiencing x. I wonder if Block isn't setting the bar too high when he characterizes what might pass for blindsight as superblindsight.

Block's distinctions, to review, are among blindsight, superblindsight and super-duperblindsight. The blindsighter has no P-consciousness of x and does not use information about x spontaneously to make inferences, guide action, etc. The superblindsighter has no P-consciousness of x but can be trained/will herself to use information about x spontaneously to make inferences, guide action, etc. The super-duperblindsighter has no P-consciousness of x, but all her relevant cognitive systems function as if she had non-degraded P-consciousness of x.

A lot seems to be riding on Block's invocation of spontaneity. So what if the typical blindsighter does not *spontaneously* form judgments about objects that impinge on her sensory organs, as long as the relevant judgments can be elicited? Block seems to contradict himself when he characterizes the blindsighter as having low-grade access. Does she have access or not? If access comes in degrees (which seems perfectly plausible), it should not matter that the blindsighter's access is of a low-grade. To repair this inconsistency (that the blindsighter has no access and low-grade access), Block really ought to say that the blindsighter doesn't have A-consciousness capable of yielding spontaneous reports.

There is the possibility that non-spontaneous access turns into low-grade access because the forced choice/training focuses the subject's attention on a phenomenal item that is too low-grade to attract attention. To understand such a scenario, we can extrapolate from the vague phenomenal items of our phenomenal periphery (although it is unclear what these peripheries are like for the non-visual dimensions of the phenomenal field) that barely attract

attention to even vaguer items that don't attract attention at all. What are they like when we attend to them? Hmm, probably a poor example, since attention focusing would tend, under normal conditions, to resolve the vague phenomenal item into a more cleanly bounded one. But what I have in mind is something that resolves into a slightly less vague phenomenal item. Maybe one could think of attending to a visual percept in one's far periphery without turning one's head or moving one's eyes.

None of this solves any problems for Block, of course, since it would still tie P-consciousness to A-consciousness.

Matt Herbert "Nim eat Nim eat.
 Drink eat me Nim.
 Me gum me gum.
 You me banana me banana you."
 --Nim the signing chimp

Get Your Private, Free Email at <http://www.hotmail.com>

From logant@U.Arizona.EDU Thu Sep 30 20:08:17 1999
Return-Path: <logant@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
 by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id UAA25843
 for <chalmers@paradox.soc-sci.arizona.edu>; Thu, 30 Sep 1999 20:08:17 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
 id <01JGLE0B404WB8WZUU@Telcom.Arizona.EDU> for
 chalmers@paradox.soc-sci.arizona.edu; Thu, 30 Sep 1999 20:08:33 MST
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU)
 by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
 with ESMTP id <01JGLE09W0GGB8W9LL@Telcom.Arizona.EDU> for
 chalmers@Arizona.EDU; Thu, 30 Sep 1999 20:08:31 -0700 (MST)
Received: from localhost (logant@localhost)
 by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id UAA29859; Thu,
 30 Sep 1999 20:08:30 -0700 (MST)
Date: Thu, 30 Sep 1999 20:08:30 -0700 (MST)
From: Logan T Trujillo <logant@U.Arizona.EDU>
Subject: Functional role of P consciousness
In-reply-to: <Pine.HPX.4.10.9909251308590.5985-100000@orion.U.Arizona.EDU>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9909301915090.13065-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Hello everyone:

I find discussions on the functional roles of A consciousness and P consciousness to be very interesting, although most psychologists would probably be uncomfortable with such a distinction. What I find most interesting about this discussion is whether or not A consciousness carries all the functionality attributed to consciousness, or does it only play a partial role, thus leaving a functional position for P consciousness. If the former case is true, then it would seem to suggest an epiphenomenal status for P consciousness, as was discussed in last Tuesday's class. Even if the latter case is true, it still seems unclear what functional role P consciousness might play.

My own view on this is in accord with the viewpoint mentioned by Dave that P consciousness is the means by which A consciousness is instantiated. Such a viewpoint obviously addresses the question of the

functional status of P consciousness by suggesting that P consciousness is fundamentally necessary for access (in terms of consciousness) to the information carried throughout various processes in the brain; hence P consciousness would not be epiphenomenal. I don't think the case of blindsight goes against this notion; evidence from the Riddoch syndrome suggests that the processing that is spared from the effects of lesioning still has a qualitative component, however impoverished and seemingly different in nature from the qualia that arises from V1 activity. Patients with intermediate blindsight claim some kinds of experiences (however strange and distorted) associated with their cognitive/perceptual appraisal of the various visual situations they are presented with experimentally. Thus it may be that this associated qualia is part of the mechanism by which the patients gain conscious access to their remaining cognitive/perceptual representations.

In the case of full blindsight this view runs into a little more difficulty in that the patients report no explicit qualia in conjunction with their guesses as towards the nature of the visual stimuli that they are presented with. However it may be that such qualia are merged into the individual's "background" qualitative state that is present due to the processing occurring in the rest of the brain. Such a background is presumably awash with all kinds of qualia that may not be explicitly distinguishable from one another. Thus the patient would not be able to explicitly differentiate them from the other qualia present in the background state, and hence would report an absence of any qualia associated with the identification task they are presented with. Therefore even in the case of full blindsight, it is still coherent to say that P consciousness plays a functional role in regards to access.

Such a viewpoint would also have something to say about the search for a NCC in that P consciousness would be present throughout all brain processes, and thus there could not be a NCC in any localized sense. Such a view would also be consistent with panpsychism in that qualia could be present in all matter/energy and would play a fundamental role in material/energetic processes (as evidenced by their role in the brain).

Logan T.

From hvorecky@U.Arizona.EDU Sun Oct 3 06:27:43 1999
Return-Path: <hvorecky@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id GAA01992
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 3 Oct 1999 06:27:43 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGP6WOP71SB8WW27@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 3 Oct 1999 13:28:45 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGP6WNIP8GB8XNWZ@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 03 Oct 1999 13:28:44 -0700 (MST)
Received: from f1n3.u.arizona.edu
(IDENT: hvorecky@f1n3.U.Arizona.EDU [128.196.137.103])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id GAA01987 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 03 Oct 1999 06:27:34 -0700
Received: from localhost (hvorecky@localhost)
by f1n3.u.arizona.edu (8.8.8/8.8.8) with ESMTP id NAA37398 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 03 Oct 1999 13:28:37 -0700
Date: Sun, 03 Oct 1999 13:28:37 -0700 (MST)
From: Juraj Hvorecky <hvorecky@U.Arizona.EDU>
Subject: Access and phenomenality
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9910031307150.49172-100000@f1n3.u.arizona.edu>
MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

I am still quite puzzled about the connection of access consciousness (AC) and phenomenal consciousness (PC). Ok, Block claims that they are fairly tied together, there are hardly any existent cases of AC without PC and there is a slight possibility of PC without AC. Let us take into account Dave's reply where he says that the exchange of the notion "poised" for the notion "available" in the definition of AC, might lead to an even stronger connection between those two. (So AC is going to be defined as "availability for (global) rational control of behavior). So far so good.

But how is the strong relation realized in the case Dave was mentioning at the end of class? I forgot the name of the guy he mentioned, but the case involved two people, one blindsight and one with a very bad vision (let's say strongly shortsighted). Now it might be the case that these two people supposedly have the same access, but they would strongly differ in phenomenality. (I am intentionally not speaking about neither AC nor PC, because supposedly there is not such a thing in blindsight.)

Well, I somehow question this conclusion. First because we are comparing a case of access to a case of AC and then show they have a different phenomenality. Well, surely they have, after all they are two different things! So if this case was supposed to show that there is a real distinction between access and phenomenality, it in fact didn't reveal anything.

Second, my prediction would be that if we would be somehow capable of measuring access (it surely is easier than measuring phenomenality), we would always end up with the same phenomenality (at least as "measured" from the first person perspective). So on one hand Dave might be right about the intrinsic relation between those two, but on the other hand if somebody is not buying the "phenomenal stuff", this strict correlation really boost Dennett in saying the second thing is an absolutely redundant.

j.

From jmartine@U.Arizona.EDU Sun Oct 3 07:21:52 1999

Return-Path: <jmartine@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id HAA02020

for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 3 Oct 1999 07:21:52 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)

id <01JGP8STLR74B8WOPT@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu; Sun, 3 Oct 1999 14:22:54 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU

(PMDF V5.1-12 #24137) with ESMTP id <01JGP8SSE2BKB8VNOS@Telcom.Arizona.EDU>

for chalmers@Arizona.EDU; Sun, 03 Oct 1999 14:22:53 -0700 (MST)

Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id HAA02011 for

<scicon@paradox.soc-sci.arizona.edu>; Sun, 03 Oct 1999 07:20:11 -0700

Received: from localhost (jmartine@localhost)

by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id OAA20577; Sun,

03 Oct 1999 14:21:12 -0700 (MST)

Date: Sun, 03 Oct 1999 14:21:12 -0700 (MST)

From: Joel A Martinez <jmartine@U.Arizona.EDU>

Subject: Functional role of P consciousness

In-reply-to: <Pine.HPX.4.10.9909301915090.13065-100000@pavo.U.Arizona.EDU>

To: Logan T Trujillo <logant@U.Arizona.EDU>

Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu

Message-id: <Pine.HPX.4.10.9910021458440.16429-100000@pavo.U.Arizona.EDU>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: RO

I agree with Logan that in the case of partial blindsight we can still find a role for P-consciousness. That is, there is still some phenomenal experience, however degraded and vague, which provides the individual with some information for making a judgement. So, these cases would be compatible with the view that it is through P-consciousness that A-conscious information is instantiated or made available.

I also agree that it is the cases of full blindsight that create a real problem. I think what Logan says here is interesting:

> In the case of full blindsight this view runs into a
> little more difficulty in that the patients report no explicit
> qualia in conjunction with their guesses as towards the nature of the
> visual stimuli that they are presented with. However it may be that such
> qualia are merged into the individual's "background" qualitative state
> that is present due to the processing occurring in the rest of the
> the brain. Such a background is presumably awash with all kinds
> of qualia that may not be explicitly distinguishable from one another.
> Thus the patient would not be able to explicitly differentiate them from
> the other qualia present in the background state, and hence would report
> an absence of any qualia associated with the identification task they are
> presented with. Therefore even in the case of full blindsight, it is still
> coherent to say that P consciousness plays a functional role in regards to
> access.

What I think is interesting here is the notion of qualia as an element of one's background state. These qualia that are merged in the background state could be taken to be like the, for lack of a better phrase, "strange qualia" mentioned in the session last Tuesday. That is, qualia that are so wierd that the brain would not know what to do with them and, hence, not process the info. as anything. (This would result in the subject not reporting any experience). I did not buy that suggestion, for various reasons, and I dont think others did either.

So, I think we want to distinguish the background state and associated qualia that Logan talks about from "strange qualia". But, this is where I become confused. How do we disntinguish the background state and associated qualia that Logan talks about from Block's examples of P-consciousness w/out A-consciousness (like the refrigerator example)? That is, Logan's description of the background state and the relevant qualia that are merged in it sounds to me like a case where one has P-consc. without A-consciousness.

The goal is to find a role for P-consciousness as the means through which A-consciousness is instantiated (as I understand it this means how the information becomes available for rational control,etc...). So, in the case of full blindsight (since this is a case of A-unconsciousness) we want to say that one is not A-conscious b/c one is not P-consc., something like that. However, we still need to explain how there is low-grade access. Logan explains this by making reference to the "background state". That is, there is still a background state that influences the behavior of the individual and gives her low-grade access.

I guess my worry is that we need to make sure that we are talking about a background state different from P-Conscious "background" states where one is A-unconscious (like my background state, including the buzz of the refrigerator, which I dont notice as I type this posting). This is because, 1), we dont want to separate P-consciousness from A-consciousness too much; and, 2), Block's examples

of P-consciousness w/out A-consciousness are very controversial (its not clear what is going on in those examples).

I dont have anything against Logan's suggestion in principle (I think it sounds interesting). I just think that in trying to keep A-consciousness and P-consciousness tied together we should be careful that we dont actually push them farther apart.

Joel M.

Joel A. Martinez
Social Science Bldg. Rm. 213
Dept. of Philosophy
University of Arizona
P.O. Box 210027
Tucson, AZ 85721-0027
USA

Office #138 Phone 520-621-7098

From chalmers@paradox.soc-sci.arizona.edu Tue Oct 5 00:07:07 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id AAA06983
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 5 Oct 1999 00:07:06 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGR7HZ2JK0B8XCUP@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Tue, 5 Oct 1999 00:07:30 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGR7HX2R5CB8X4NJ@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Tue, 05 Oct 1999 00:07:27 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id AAA06977 for scicon@paradox.soc-sci.arizona.edu; Tue,
05 Oct 1999 00:06:58 -0700
Date: Tue, 05 Oct 1999 00:06:58 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: A-con, P-con, blindsight
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199910050706.AAA06977@paradox.soc-sci.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; CHARSET=US-ASCII
Status: R0

Re A and P consciousness, and blindsight:

It's very important to remember that access consciousness and mere access are not the same thing. There are zillions of different sorts of access. Access consciousness is the very specific sort where information that one has access to is made available for verbal report, rational inference, and spontaneous voluntary control of behavior, and so on. Most instances of access won't be instances of access consciousness.

In particular, the access that a blindsight patient has won't necessarily be access consciousness. The point of talk of "low-grade access" is for just this reason. High-grade access is the specific sort of access above that corresponds to A-consciousness. Low-grade

access is mere access that may have some impact on brain or behavior, but that doesn't play the role specified in the definition of A-consciousness.

Matt suggests that Block's invocation of spontaneity in the definition of A-consciousness leads to inconsistency, since one can have access without spontaneity. I think this would only lead to inconsistency if it were held that all instances of access should be instances of A-consciousness. As it is, Block can define A-consciousness how he likes, so he is welcome to include spontaneity in his definition if he wants.

Juraj asks about Siewert's case discussed in class of (i) Belinda with blindsight on one visual hemifield and (ii) Connie with low-grade phenomenal vision in that hemifield. These two people might have quite similar functional capacities. One difference is that the blindsight patient might not make spontaneous judgments. Siewert in his discussion stipulates that they do make these judgments; the cost is that the case may now be impossible in nature, but he says it is quite conceptually possible. Apart from this, the obvious difference is that Belinda will report conscious vision and Connie won't; Siewert is willing to keep this difference, so this isn't a zombie case.

If we stipulate spontaneous judgments and voluntary control, etc, then it seems that both Belinda and Connie have visual A-consciousness, but only Belinda has visual P-consciousness (this is part of Siewert's point about the irreducibility of P-consciousness). Of course Belinda is somewhat hypothetical, so this may not be a counterexample to the link between A and P in the actual world. If we go with actual counterparts of Connie, we may have to lose spontaneous judgments etc, and thereby lose visual A-consciousness. Either way, measurements of A-con will differ slightly from measurements of P-con here, since Connie will have verbal reports *of consciousness* and Belinda won't. So we don't obviously end up collapsing the notions into one, even on measurement grounds.

Re a functional role for P-consciousness: I agree with Logan and Joel M. that all the empirical results are compatible with the hypothesis that the role of P-consciousness is to enable A-consciousness. Though I do think there are other problems here, concerning the sufficiency of neural mechanisms and the apparent redundancy of P-consciousness in this role.

I like Joel's idea that "strange qualia" that we can't deal with are potentially cases of P-con without A-con. No rational control etc, just phenomenality. Of course it is not clear that there are really such cases. Neglect is one interesting possibility, though one could argue that there is limited A-con in such a case, and the phenomenal aspect of neglect is never clear. But anyway, it does seem that these strange qualia might have the effect of separating P and A.

So if Logan wants to keep the two associated, it might be better to make the "background state" that he appeals to in blindsight either not a matter of qualia, or a matter of a special sort of qualia that don't play the usual access role. (E.g., maybe nonstrange qualia play the access role, and strange qualia play the enabling role?) I suspect that we don't need to say qualia do everything here -- we know that lots of what happens in the brain (e.g. low-grade access everywhere) is not obviously qualia-associated, so maybe one could leave the role of qualia for the high-grade stuff. Though of course there are still the general epiphenomenalistic problems mentioned above.

--Dave.

From press@U.Arizona.EDU Sun Oct 10 15:49:01 1999
Return-Path: <press@U.Arizona.EDU>
Received: from Arizona.EDU (Hopey.Telcom.Arizona.EDU [128.196.128.234])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id PAA10311
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 15:49:01 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JGZ3U7PSK09AMIDF@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
10 Oct 1999 15:49:10 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JGZ3U75YN49AMIDE@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 10 Oct 1999 15:49:10 -0700 (MST)
Received: from f1n3.u.arizona.edu
(IDENT:press@f1n3.U.Arizona.EDU [128.196.137.103])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id PAA10306 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 15:48:54 -0700
Received: from localhost (press@localhost) by f1n3.u.arizona.edu (8.8.8/8.8.8)
with ESMTP id PAA35230 for <scicon@paradox.soc-sci.arizona.edu>; Sun,
10 Oct 1999 15:49:03 -0700
Date: Sun, 10 Oct 1999 15:49:03 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>
Subject: Causal Relativism
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9910101545440.39446-100000@f1n3.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

All,

At the end of our last session, we only briefly touched on Velmans' (rather bizarre) claim that phenomenal consciousness plays a causal role in the production of behavior on a first person description, but not on a third person description. Even so, we probably gave it as much consideration as it really deserves. The consensus of the class at least seemed to be that "causal relativism" doesn't make sense. I certainly agree, but I think that Velmans' point is a bit more sophisticated than our brief discussion of it let on.

The relevant paragraph reads as follows:

"Of course, the question of whether it is actually a given brain state or a given experience that determines a particular behaviour remains - and, on this point, I suggest, no choice is necessary. These are events viewed from different perspectives. Events viewed from an external observer's perspective (via exteroceptors) appear different from the same events experienced by the subject (via interoceptors) because the methods of observation are different. However, each perspective is legitimate."

The emphasis here seems to be on regarding introspection as a type of perception. If introspection is similar to the other modes of perception, then it seems reasonable to say that we should not give any more or less weight to this particular mode of perception. If the data (about the very same event) collected from different sensory modalities conflicts, we may find ourselves with two very different models of the world based on those different sets of data.

So he seems to be thinking that the difference between first and third person accounts of the causal role of consciousness (or lack thereof) is like a situation in which two of our other senses conflict. For example, consider the case of a person confronted with a cup sitting on a transparent glass tabletop. If the tabletop is invisible to our subject

(suppose it's really clean, no light glinting off it, etc.), the cup might look like it is levitating in thin air. However, if the subject is allowed to touch the tabletop, he will believe that the cup is resting on it.

Now, if there were no way to cross-check the data gathered by these two different senses, we might be forced to accept both descriptions as equally valid, even though they were inconsistent. However, it generally seems to be the case that we can cross-check. Our table-observing subject can probably move to a different observational angle, shine a bright light on the table, or whatever. Or, if cross-checking is impossible, we can at least explain why one sense modality fails to detect some fact (or appears to detect an inconsistent one.) Even if our tabletop were made out of some sort of exotically invisible material (maybe a Star Trek type force field or something) we ought to be able to use data gathered from our other senses and/or other visual data to explain why our eyes can't detect it.

Velmans seems to think that this will not be possible with consciousness. Both the first and third person perspectives are, he says, incomplete. Neither is adequate to explain the other. Though I disagree with this position, it does not seem entirely unreasonable to me. Since none of our means of perception, including introspection, can, by itself, give us a complete description of the world, we should perhaps not expect the various descriptions they provide to always be consistent. Perhaps first and third person descriptions are accurately describing different aspects of the same thing, but both fail to include the data that would link them.

Even so, I am not about to grant Velmans his conclusion, because he is too quick to dismiss the possibility that an improved first or third person account will be able to account for the alternate set of perceptions. Even if there is good evidence that the two accounts are incompatible (and this does not seem clear either) I think most of us would be extremely reluctant to give up on the attempt at a unified explanation. Our third person description of the brain and its functioning is obviously incomplete, and as many of the issues we have been discussing lately show, introspection may not provide as detailed a description as we might have thought. (For example, if a blindsight patient cannot tell whether his own behavior was accompanied by no qualia, very faint qualia, "weird" qualia, or neglected qualia, I for one start to wonder how complete our introspective accounts of our mental lives really are.) Surely Velmans is premature in his declaration that no future improvement of one of these descriptions could ever explain the seemingly inconsistent perceptions of the other.

Joel

From bradt@U.Arizona.EDU Sun Oct 10 17:56:15 1999

Return-Path: <bradt@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.Arizona.EDU [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA10660

for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 17:56:15 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)

id <01JGZ8AVWYDS9BVD4X@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,

10 Oct 1999 17:56:22 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU

(PMDF V5.2-31 #39830) with ESMTP id <01JGZ8AUBD5S9ED96K@Telcom.Arizona.EDU>

for chalmers@Ariziona.EDU; Sun, 10 Oct 1999 17:56:20 -0700 (MST)

Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA10654 for

<scicon@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 17:55:55 -0700
Received: from localhost (bradt@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id RAA14175; Sun,
10 Oct 1999 17:56:04 -0700 (MST)
Date: Sun, 10 Oct 1999 17:56:04 -0700 (MST)
From: Brad J Thompson <bradt@U.Arizona.EDU>
Subject: Re: Causal Relativism
In-reply-to: <Pine.A41.4.10.9910101545440.39446-100000@f1n3.u.arizona.edu>
To: Joel K Press <press@U.Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910101737060.5529-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

I think Joel is right to suggest that Velman's proposal concerning the complementarity of first and third-person perspectives regarding their causal status deserves a second-glance. I gave it a second-glance after reading Joel's post, and am now prepared to defend it completely! In fact, Velman's view speaks directly to the issue we've discussed off-and-on regarding the possible causal role of P-consciousness (perhaps as a "gateway" to access--Lis mentions this idea again in her post, and Block seemed to be attracted to it).

The big problem with the idea that P-consciousness serves as the means of an event or state's becoming A-conscious is that the metaphysics is difficult to make sense of. My zombie twin has all the same A-conscious states but no P-conscious states. This suggests that P-consciousness isn't doing any real causal work. Another way to the same point is simply to note that the availability of a state for global control (etc.) could be understood entirely from the third-person point of view even if the observer did not know which states were P-conscious or not. Even if P-conscious states are identical to neural states, if those states are what "give rise" to A-consciousness it is not in virtue of their being P-conscious.

The above point is entirely compatible with what Velman's suggestion. The third-person point of view leaves out phenomenal consciousness, and doesn't need to posit P-consciousness as a variable which plays a role in human information processing. But consciousness certainly **seems** to play a causal role in our behavior. Velman saves us from epiphenomenalism by arguing that the two perspectives are compatible. I would read his view (at least in the passages Joel cites) as a type of identity theorist, where perhaps we are directly acquainted with the intrinsic nature of neural states from the first-person perspective. But Dave suggested in class that Velman's was some type of idealist, so perhaps Velman's has developed these views in a different direction.

Brad J Thompson
bradt@U.Arizona.EDU

From serobert@U.Arizona.EDU Sun Oct 10 21:37:05 1999
Return-Path: <serobert@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.Arizona.EDU [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA11013
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 21:37:05 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JGZG0QFI8W9D4DHI@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
10 Oct 1999 21:37:15 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JGZG0Q20289ED9HR@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 10 Oct 1999 21:37:14 -0700 (MST)
Received: from fln2.u.arizona.edu
(IDENT:serobert@fln2.U.Arizona.EDU [128.196.137.102])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA11008 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 21:36:54 -0700
Received: from localhost (serobert@localhost)
by fln2.u.arizona.edu (8.8.8/8.8.8) with ESMTP id VAA35708; Sun,
10 Oct 1999 21:37:03 -0700
Date: Sun, 10 Oct 1999 21:37:02 -0700 (MST)
From: Simon E Roberts-Thomson <serobert@U.Arizona.EDU>
Subject: Re: Causal Relativism
In-reply-to: <Pine.HPX.4.10.9910101737060.5529-100000@orion.U.Arizona.EDU>
To: Brad J Thompson <bradt@U.Arizona.EDU>
Cc: Joel K Press <press@U.Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9910102058070.38712-100000@fln2.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

I would just like to make a few points in response to Joel and Brad. As Joel pointed out, we seemed to have decided in class that Velmans was committed to some sort of causal relativism, and hence we pretty much ignored what he had to say. Both Joel and Brad, however, seem to contend that Velmans has something worthwhile to say. As Brad says:

> human information processing. But consciousness certainly *seems* to
> play a causal role in our behavior. Velman saves us from epiphenomenalism
> by arguing that the two perspectives are compatible. I would read his
> view (at least in the passages Joel cites) as a type of identity theorist,
> where perhaps we are directly acquainted with the intrinsic nature of
> neural states from the first-person perspective. But Dave suggested in

Whilst I agree that this particular passage may be read as suggesting that Velmans is some sort of identity theorist, this is incompatible with what he says elsewhere. In 9.2 he says:

"It should be apparent that the dissociation of awareness from cerebral functioning poses problems for reductionist theories of consciousness. It is inconsistent with the functionalist view that consciousness simply is (ontologically identical to) a mode of functioning of the brain. Nor is it consistent with physicalism ... The exclusion of consciousness from cerebral functioning is equally inconsistent with emergent interactionism ... and with the interactionist forms of dualism"

It is difficult to say exactly what sort of theory he advocates, but it does not seem to be an identity theory. Velmans goes on to say that the psychological evidence seems "to support epiphenomenalism", but that "From a first person perspective ... epiphenomenalism appears false". From this he concludes that these two perspectives are "complementary, [but] mutually irreducible". Needless to say, this conclusion is inadequate. It seems to me that the reasonable conclusion to draw from the evidence is not that epiphenomenalism is both true and false, depending on the viewpoint adopted, but that one of these viewpoints must be mistaken. Velmans does seem to be arguing for some sort of relativism, and this view cannot be sustained by the evidence that he puts forward. We should either be arguing that the information processing models that he adopts are incomplete, in so far as they do not account for our first person accounts of the causal efficacy of consciousness, or we should argue that our first person accounts are mistaken.

Thus I think that the conclusions that Velmans draws are disputable at best, and just plain wrong at worst. We should not accept that epiphenomenalism is both true and false, depending on your point of view, unless there is absolutely no other conclusion that we can draw. In the case above, I think that there are many other more plausible conclusions that can be drawn, and hence we should reject Velmans' conclusions.

Simon.

From franzen@U.Arizona.EDU Sun Oct 3 09:41:09 1999
Return-Path: <franzen@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id JAA02337
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 3 Oct 1999 09:41:09 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.1-12 #24137)
id <01JGPDNIIVJ4B8XAL3@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu; Sun, 3 Oct 1999 16:42:11 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.1-12 #24137) with ESMTP id <01JGPDNH5TKWB8SZFD@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 03 Oct 1999 16:42:09 -0700 (MST)
Received: from f1n3.u.arizona.edu
(IDENT:franzen@f1n3.U.Arizona.EDU [128.196.137.103])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id JAA02332 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 03 Oct 1999 09:40:56 -0700
Received: from localhost (franzen@localhost)
by f1n3.u.arizona.edu (8.8.8/8.8.8) with ESMTP id QAA49014 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 03 Oct 1999 16:41:59 -0700
Date: Sun, 03 Oct 1999 16:41:59 -0700 (MST)
From: Peter L Franzen <franzen@U.Arizona.EDU>
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9910031641060.39540-100000@f1n3.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

On the issue of fleeting awareness in unconscious perception: If you are exposed to something briefly, or while in a "less-conscious" state such as sleep or under anesthesia, what does this say about the experience? For example, healthy normal sleepers have, on average, 4-5 dreams per night. Many people do not recall this many, as you need to have a brief awakening at the end of the REM sequence in order to have any free recall of the dream. So, if you have a dream, but don't remember it, did you have a phenomenal experience? I would say yes, although it is not remembered. If an experience is not remembered, or falls away immediately (from memory), the experience could still be access-consciousness (such as in priming). Just because you don't remember a particular dream during the night does not mean that this experience will not have an affect on you the following day (through some kind of priming or perhaps a deja vu experience). This seems to relate to the experiments where the 3 x 3 grid of letters are flashed, and the subject can report on any of the lines, but not all of them. This suggests that the person is phenomenally aware of all of the stimuli (for some time afterward), but only access conscious of one line.

From lnielsen@azstarnet.com Sun Oct 10 16:19:56 1999
Return-Path: <lnielsen@azstarnet.com>
Received: from Arizona.EDU (Penny.Telcom.Arizona.EDU [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA10481

for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 16:19:56 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)

id <01JGZ4XHG2Q89BVCPK@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,

10 Oct 1999 16:20:04 MST

Received: from paradox.soc-sci.arizona.edu ([128.196.44.83])

by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)

with ESMTP id <01JGZ4XFZ7C09ED8YW@Telcom.Arizona.EDU> for

chalmers@Arizona.EDU; Sun, 10 Oct 1999 16:20:01 -0700 (MST)

Received: from cepheus.azstarnet.com (cepheus.azstarnet.com [169.197.56.195])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA10477 for

<scicon@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 16:19:38 -0700

Received: from zippo (dialup10ip097.tus.azstarnet.com [169.197.34.225])

by cepheus.azstarnet.com (8.9.3+blt.Beta0/8.9.3) with SMTP id QAA18037 for

<scicon@paradox.soc-sci.arizona.edu>; Sun, 10 Oct 1999 16:19:42 -0700 (MST)

Date: Sun, 10 Oct 1999 16:25:31 -0700

From: Lis Nielsen <l Nielsen@azstarnet.com>

Subject: How limited is the unconscious?

To: Scicon <scicon@paradox.soc-sci.arizona.edu>

Reply-to: l Nielsen@u.arizona.edu

Message-id: <NCBBJCIPCKAPOGDICDGPCECECDAA.l Nielsen@azstarnet.com>

MIME-version: 1.0

X-MIMEOLE: Produced By Microsoft MimeOLE V5.00.2314.1300

X-Mailer: Microsoft Outlook IM0, Build 9.0.2416 (9.0.2910.0)

Content-type: text/plain; charset=iso-8859-1

Content-transfer-encoding: 7bit

Importance: Normal

X-Priority: 3 (Normal)

X-MSMail-priority: Normal

X-Sent-via: StarNet <http://www.azstarnet.com/>

Status: RO

Re the limitations of the unconscious: As Dave pointed out, there seems to be a tension between Greenwald's claims that "attentionless unconscious cognition is generally quite limited in the sophistication of analyses it can perform," and Velmans' claims to the effect that much complex cognition can occur outside of consciousness.

As I said in class, I don't think that the two claims made here are incompatible. Greenwald and Velmans are both concerned with how information comes to play a function in complex cognition and under what conditions this is possible. Both argue that attention is key to these functions, and that manipulations that disrupt attention leave behind only very low level influences of the presented information on subsequent processing. The critical issue concerns whether such information can become access-conscious, to use Block's term (either in the sense of poised for direct, voluntary control of action, etc., or in the sense of cerebral celebrity, which I take to mean currently being in such control). From Greenwald's review, it seems that information must be P-conscious in order to be available for complex cognition. Velmans argues that it is focal attention rather than P-consciousness that enables access.

The various manipulations in unconscious perception studies degrade P-consciousness sufficiently to limit the access of the information to global processing. In the absence of high-grade P-consciousness, the information that gets into the system can only play a passive role, via processes like the spreading of activation in semantic networks or the priming of motor, affective or other responses. The passivity of these processes lies in their purely data-driven nature; spreading of activation

or priming cannot lead to interpretation or evaluation of the information at the conscious level nor, perhaps, does it allow its active integration as premises in 'unconscious reasoning', i.e., unconscious processes like addition that are set in motion by conscious presentation of inputs. However, if a context for interpretation or response is provided (say one is looking for an answer to a difficult problem or one is required to give a word stem completion in an experimental task) then the information can have effects on these outputs. These effects appear in multiple ways - as sudden insights or intuitions, as response biases or shortened response latencies, without the subject having explicit awareness of how these effects came about. The processing of information that is presented under conditions where subjects cannot report consciousness of the information has this kind of passive characteristic.

The passive, data-driven processing set in motion by unconsciously perceived information stands in contrast to the active, theory-driven processing of information that is consciously perceived. For example, what we consciously perceive sets up expectations for what we might see next; it enables us to organize complex behaviors in response to the information. That many aspects of these subsequent cognitive functions will proceed unconsciously seems to be Velmans' main point in the first half of his article. This sophistication of the unconscious under normal stimulus conditions is widely accepted, though Velmans wants to make the further point that it is not due to P-consciousness that this kind of information processing takes place, rather that focal attention (A-consciousness) is all that is required. I don't think Greenwald would challenge Velmans' claims about the sophistication of the unconscious under normal conditions. He might, however, take issue with the claim that focal attention alone has the function of making information widely available. Focal attention is present in unconscious perception experiments, it simply doesn't have the right kinds of inputs to operate on. Greenwald wants to emphasize is that high-level unconscious information processing doesn't get going all on its own, even if focal attention is engaged. Sophisticated unconscious processing requires that information be the object of attention, and that it be so long enough to set up an enduring conscious representation of that information.

One of the implications of the unconscious perception experiments is that we are not vulnerable to unlimited, complex, subliminal manipulations that bias our behavior. In order for information to actively play a role in ongoing behavior and cognition, it needs to be P-conscious first, much along the lines of Block's current claim that P-consciousness is the means by which information becomes A-conscious. The effects of "unconsciously perceived" stimuli discovered so far do not meet the criteria for A-consciousness, which seems to line up well with this analysis.

Now, given Block's reasonable suggestion on Friday that there is fleeting P-consciousness in the unconscious perception cases (without A- or R-consciousness), we would want to say that while P-consciousness might be necessary for A-consciousness, it is not sufficient for A-consciousness. Other conditions are required, involving contextual and temporal factors that may vary from situation to situation to set up the kind of representational stability suggested above. It would seem that our notion of P-consciousness will require some revision.

Lis Nielsen
Department of Psychology
University of Arizona
Tucson, AZ 85721-0068
USA

lnielsen@u.arizona.edu

From switanek@U.Arizona.EDU Tue Oct 12 14:43:36 1999
Return-Path: <switanek@U.Arizona.EDU>
Received: from Arizona.EDU (Hopey.Telcom.Arizona.EDU [128.196.128.234])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id OAA15542
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 12 Oct 1999 14:43:36 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JH1U5TQR689D4QAT@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Tue,
12 Oct 1999 14:43:48 MST
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU)
by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
with ESMTP id <01JH1U5TDS409EDI01@Telcom.Arizona.EDU> for
chalmers@Ariziona.EDU; Tue, 12 Oct 1999 14:43:47 -0700 (MST)
Received: from localhost (switanek@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id OAA06045; Tue,
12 Oct 1999 14:43:47 -0700 (MST)
Date: Tue, 12 Oct 1999 14:43:47 -0700 (MST)
From: Nicholas J Switanek <switanek@U.Arizona.EDU>
Subject: attention and conscious perception
In-reply-to: <199910050642.XAA06950@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Ariziona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910121403130.7266-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Our discourse of last week and of this has as one of its foci the distinction between conscious and unconscious perception. Conscious perception has its distinguishing characteristics 1)a phenomenology that extends through time--it's not instantaneous--2)accessibility, in terms of availability for verbal report, and instantiation in memory. Unconscious perception is a trickier thing. It appears from the Mack and Rock experiments that visual percepts, at least those relatively near to point of visual attention, are encoded in some detail, although they are never made available for verbal report and never make it into memory. In fact, between the two, availability for verbal report and memory, memory is clearly the more fundamental. I don't think it makes much sense to talk about instantaneous conscious perception and I think it is memory that allows perceptions that might have been collected at discrete moments to extend through time. But I'm not sure that I could go as far as saying conscious perception is only unconscious perception plus memory of the percepts; actually, I'm sure I can't: necessary is some sort of organization and ordering of the memories.

One way to think about unconscious perception might be to call it a variety of priming. When we are asked a question or by some other means instigate a search for something in our perceptual field, items that are already unconsciously encoded pop out or grab our attention. In a way it seems the brain is continually overprocessing, preparing information that mostly gets dumped, never makes it into consciousness, the stream of occurrent thought, or into memory. Encoded information about the visual field might be readied in case it needs being attended to, but since the higher-level powers that be deem the information of lower priority or irrelevant to the tasks at hand, never call the info over the threshold into consciousness.

So what decides which primed pathways will be trodden by victorious electrical impulses into the Cartesian circus? Well, the individual has some explicit control over what searches it establishes, and once the search is on, an answer receptacle is prepared in memory. There might also be other searches that are continually and quietly going on. Searches for particular learned or instinctual icons might be such latent and implicit procedures that run constantly. Or if these searches don't run constantly, they run especially smoothly, for how else is a complex icon like one's own name recognized with rapidity far exceeding that for other percepts?

nick

From rachaelp@U.Arizona.EDU Wed Oct 13 12:39:51 1999
Return-Path: <rachaelp@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id MAA17363
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 13 Oct 1999 12:39:50 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JH344RGBC09D4R9H@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@Arizona.EDU); Wed,
13 Oct 1999 12:40:04 MST
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU)
by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
with ESMTP id <01JH344R07CW9EDP78@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Wed, 13 Oct 1999 12:40:03 -0700 (MST)
Received: from localhost (rachaelp@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id MAA06992; Wed,
13 Oct 1999 12:40:02 -0700 (MST)
Date: Wed, 13 Oct 1999 12:40:02 -0700 (MST)
From: Rachael J Parkinson <rachaelp@U.Arizona.EDU>
Subject: attention and consciousness
In-reply-to: <Pine.HPX.4.10.9910121403130.7266-100000@pavo.U.Arizona.EDU>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910131145320.1409-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

In seminar yesterday we discussed three different views of the nature of consciousness and attention and considered them as possible interpretations of the change blindness studies. As these views are relevant to my comment, I will go over them here briefly:

- 1) The 'grand illusion' view: Our visual consciousness is coarse-grained with limited internal representation. On this view, there is no consciousness outside attention. This is the view that we identified with Dennett and Mack and Rock.
- 2) The reconstructionist theory of vision: There is no grand illusion. In fact, we have fine grained visual consciousness that gives rise to detailed internal representation. We do have consciousness without attention. To account for the change blindness studies, some explanation may be offered. For example, perhaps we do not have memory without attention. This view is associated with Wolfe and Hardcastle.
- 3) The enactive approach: On this view, we do have fine grained visual consciousness though we only have a limited internal representation. There is consciousness outside of attention. But we must be clear, the fine-grained detail of our visual consciousness is a result of external availability and is not a matter of internal representation. Thus, the contents of consciousness need only be *available* in the world. this externalist view is the one espoused by Noe, Pessoa, and Thompson.

I would like to suggest a fourth option. On this view, there is consciousness outside of attention, but that visual consciousness is coarse-grained, while the consciousness within attention is fine-grained. Thus, we have detailed internal representation of what we are paying attention to and limited internal representation of what we are not. Let me illustrate. We have all had the experience of driving without paying attention or of staring off into space. If attention were required for consciousness, it seems that we would not be phenomenally aware *in any sense* of the outside world. But this does not seem to me to be the case when I reflect on my own experience. In these cases, I would want to say

that I was limitedly visually conscious (of course I could not describe the detail of my visual environment) even though my attention is directed elsewhere. If for example, the whole world were to disappear in one of these 'reflective' moments I would be more inclined to say that I was conscious of the world around me and the change directed my attention to it then that I was not conscious of the world until it's 'disappearance'.

I think this view make sense of the cocktail party phenomenon and others like it as it seems that some conscious processing must be going on for your name to jump out at you. We are limitedly phenomenally conscious of what is going on around us at the cocktail party, though we are not paying attention. Hearing our name spoken is something we are already conscious of but now pay attention to.

This view may sound a lot like the third view proposed by Noe, Pessoa, and Thompson. But this view is not meant to be an externalist view. It does not merely require that the contents of consciousness be available in the outside world (a very tricky notion as Dave pointed out in class.) The contents of consciousness are those that we have internal representations of; those that we are paying more attention to will be more detailed, those that we are not paying attention to will be significantly less detailed. Maybe we will only be vaguely aware of color, shape, size outside of our attention. Minimally, we will be aware of the existence of this coarse grained visual experience. This is compatible with the change blindness studies because I am not suggesting that we will be aware of particular changes in our unattended environment. Only that we are aware, in a general and vague sense, of our unattended environment. Perhaps it is this awareness that gives rise to the illusion of a complete detailed visual consciousness.

Sorry for the length of my post. Any thoughts on this view?

Best,
-Rachael

From chalmers@paradox.soc-sci.arizona.edu Wed Oct 13 21:37:18 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Hopey.Telcom.arizona.edu [128.196.128.234])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA18339
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 13 Oct 1999 21:37:18 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JH3MW4RQU89D4VU4@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Wed,
13 Oct 1999 21:37:32 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JH3MW4KM689EDT2N@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Wed, 13 Oct 1999 21:37:31 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id VAA18328 for scicon; Wed, 13 Oct 1999 21:37:05 -0700
Date: Wed, 13 Oct 1999 21:37:05 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Consciousness and attention, etc.
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199910140437.VAA18328@paradox.soc-sci.arizona.edu>
Status: R

Hi, a few notes on various things:

(1) Tim distinguishes two notions of attention, a phenomenal sort and a non-phenomenal sort. That seems reasonable to me. We might talk about P-attention and A-attention! Tim also asks whether blindsight patients can attend to their blindsight. Interestingly there are

results from Robert Kentridge, presented at Tucson III, showing that blindsight subject's performance is improved when they attend to the relevant areas of the visual field. So we have a sort of attention to the object without phenomenal consciousness of the object? Or perhaps, it's safer to say that it's just attention to the spatial location where the object is present. That way, the contents of attention and the contents of consciousness will still line up.

(2) Nick suggests that it doesn't make much sense to talk about instantaneous conscious perception, and that memory may be required for consciousness. That's an interesting claim. If one accepted this claim, then one might argue that the change blindness studies show that we aren't conscious of the relevant detail in our visual field; and it might suggest that attention mechanisms operate on largely unconscious (not just unattended) stimuli. But why accept that claim? I'd be interested to see reasons for it.

(3) Rachael suggests a view on which we have fine-grained consciousness within attention, and coarse-grained consciousness outside attention. That seems reasonable and phenomenologically apt to me -- it seems reasonable that the "attentional spotlight" makes things more vivid and detailed. One relevant question, I guess, is just how coarse-grained the consciousness outside attention is. For example, in the change blindness cases, will there be a change (unnoticed, of course) in the coarse-grained consciousness outside attention, or will it stay the same? If the latter, then we are not so far from the Dennett and Mack/Rock views. If the former, we are closer to the Wolfe/Hardcastle view, with the qualification that consciousness outside attention is not as fine-grained as we might think.

(4) If anyone is interested to check out the web demos from Wednesday, they are online (along with a few articles) at:

<http://nivea.psych.univ-paris5.fr/>

For something different but very interesting in the ballpark of "experimental phenomenology", check out:

<http://www.headless.org/>

--Dave.

From lan@U.Arizona.EDU Thu Oct 14 11:17:33 1999

Return-Path: <lan@U.Arizona.EDU>

Received: from Arizona.EDU (Hopey.Telcom.arizona.edu [128.196.128.234])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id LAA01668

for <chalmers@paradox.soc-sci.arizona.edu>; Thu, 14 Oct 1999 11:17:33 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)

id <01JH4FJ9VN809D5B98@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Thu,

14 Oct 1999 11:17:55 MST

Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU)

by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)

with ESMTP id <01JH4FJ9KN8G9EDX7U@Telcom.Arizona.EDU> for

chalmers@Ariziona.EDU; Thu, 14 Oct 1999 11:17:54 -0700 (MST)

Received: from localhost (lan@localhost)

by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190))/8.8.6) with ESMTP id LAA18153; Thu,

14 Oct 1999 11:17:54 -0700 (MST)

Date: Thu, 14 Oct 1999 11:17:53 -0700 (MST)

From: Lonnie A Nelson <lan@U.Arizona.EDU>

Subject: Re: Consciousness and attention, etc.

In-reply-to: <199910140437.VAA18328@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910141059170.630-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Hi all-

A couple of comments and questions...

>
> (2) Nick suggests that it doesn't make much sense to talk about
> instaneous conscious perception, and that memory may be required for
> consciousness. That's an interesting claim. If one accepted this
> claim, then one might argue that the change blindness studies show
> that we aren't conscious of the relevant detail in our visual field;
> and it might suggest that attention mechanisms operate on largely
> unconscious (not just unattended) stimuli. But why accept that claim?
> I'd be interested to see reasons for it.

The closest thing that I can think of to a reason for this (though I was not the person who proposed it) is in line with the Physicist Fred Allan Wolf's assertion that we are never conscious of anything as it occurs, there is at least, the time taken for the light to reach the eye, the neural signals to become processed, etc. On that view, we are always time t behind the world. But this does not speak to whether or not memory is necessary for consciousness since we would not have to have memory of what the world experienced as we would be a constant (t) behind actual occurrence and would not need to access our memory (that was never formed) from that time frame.

However, for the change blindness business, memory may be the determining factor of whether or not change is noticed, unless you are looking at the thing that is changing when it changes, since then you have to compare your memory to the current picture in order to detect a change.

> (3) Rachael suggests a view on which we have fine-grained
> consciousness within attention, and coarse-grained consciousness
> outside attention. That seems reasonable and phenomenologically apt
> to me -- it seems reasonable that the "attentional spotlight" makes
> things more vivid and detailed. One relevant question, I guess, is
> just how coarse-grained the consciousness outside attention is. For
> example, in the change blindness cases, will there be a change
> (unnoticed, of course) in the coarse-grained consciousness outside
> attention, or will it stay the same? If the latter, then we are not
> so far from the Dennett and Mack/Rock views. If the former, we are
> closer to the Wolfe/Hardcastle view, with the qualification that
> consciousness outside attention is not as fine-grained as we might
> think.
>

This view makes more sense to me than any of the others presented so far. As far as the degree of coarse grainedness of the consciousness outside of attentional focus, I would bring up two variables that could be looked to as possible answer providing lines of thought (or research). The first would be the individual person's "attentional sharpness" that is, the degree of detail that they are capable of perceiving visually. Someone with horrid eyesight is going to be likely to have less attention invested in that mode to begin with, and when they do, they will probably be using a pen light as opposed to a mag-lite so to speak. Another variable that I would look to is the degree of patterning present in the stimuli outside the focus of attention. Since the findings

regarding the relevance of a stimulus indicate that pattern detection exists outside of attention, one could say that this coarse grained consciousness is constantly undergoing change of content. Unless Dave is referring to something else by "change in consciousness."

Just thoughts.
Lonnie

It is a common fate of all knowledge to begin as heresy and end as orthodoxy.

-Thomas Huxley

Lonnie A Nelson
Department of Psychology
Human Energy Systems Laboratory
University of Arizona
lan@u.arizona.edu

From hvorecky@U.Arizona.EDU Sun Oct 17 16:30:24 1999
Return-Path: <hvorecky@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA08990
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 16:30:24 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JH8XCAJS1C9D7HGB@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
17 Oct 1999 16:30:51 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JH8XCAAEM09EE0Z4@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 17 Oct 1999 16:30:50 -0700 (MST)
Received: from f1n2.u.arizona.edu
(IDENT: hvorecky@f1n2.U.Arizona.EDU [128.196.137.102])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA08981 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 16:30:18 -0700
Received: from localhost (hvorecky@localhost)
by f1n2.u.arizona.edu (8.8.8/8.8.8) with ESMTP id QAA41162 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 16:30:45 -0700
Date: Sun, 17 Oct 1999 16:30:45 -0700 (MST)
From: Juraj Hvorecky <hvorecky@U.Arizona.EDU>
Subject: Attention and A-con
To: Science of Consciousness <scicon@paradox.soc-sci.arizona.edu>
Message-id: <Pine.A41.4.10.9910171613510.34986-100000@f1n2.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Last class had made me a little puzzled. Prior to it I have always had a strong feeling about the correlation between consciousness and attention. Because so little is known about P-consciousness, I guess I will be more or less talking about AC now.

First of all, the clear difference between attention and AC seems to be that one is actual and the other dispositional. Well, exactly this difference is a little mystery to me. What does it mean to talk about being dispositional here? I guess I am only conscious of the events which do actually make it to my attention. There is no awareness whatsoever about any disposition. Consciousness seems to possess no dispositional qualities. It is just what is being realized.

One might object that certain inconclusive reports made by conscious

subjects imply dispositional character of AC. Statements such as "I have seen something on the screen..." in my understanding show no such a character. They simply refer to what was actually realized, to what the attention was paid. That the attention wasn't sufficient for a detailed account of an event implies very little about the event "being potentially accessible".

I apologize for being probably too messy here. If I have to put it in a different wording, it would be something like this - my first person experience (not restricted to phenomenality) seems to recognize no dispositional account of consciousness, but supposedly vast majority of my AC events is just that - dispositional mental events.

I wonder about this, because the third person description of attention seems to perfectly fit the first person perspective - I perceive something as being attended anytime any third person observer perceives me attending to something. There is no talk about disposition/realization.

But it seems like this in the case of AC does not really work

juraj

From logant@U.Arizona.EDU Sun Oct 17 16:32:38 1999
Return-Path: <logant@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA09001
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 16:32:38 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JH8XF2BISW9D7BN0@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
17 Oct 1999 16:33:05 MST
Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
with ESMTP id <01JH8XF2225S9EEZ5X@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Sun, 17 Oct 1999 16:33:05 -0700 (MST)
Received: from localhost (logant@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id QAA05102; Sun,
17 Oct 1999 16:33:04 -0700 (MST)
Date: Sun, 17 Oct 1999 16:33:04 -0700 (MST)
From: Logan T Trujillo <logant@U.Arizona.EDU>
Subject: Re: attention and consciousness
In-reply-to: <Pine.HPX.4.10.9910131145320.1409-100000@pavo.U.Arizona.EDU>
To: Rachael J Parkinson <rachaelp@U.Arizona.EDU>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910171524390.10779-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Hello all:

In a recent posting Rachel proposed a fourth alternative to the three viewpoints discussed in last Tuesday's class concerning the relationship between consciousness and attention. Here is a quote from her posting:

> I would like to suggest a fourth option. On this view, there is
> consciousness outside of attention, but that visual consciousness is
> coarse-grained, while the consciousness within attention is fine-grained.
> Thus, we have detailed internal representation of what we are paying
> attention to and limited internal representation of what we are not. Let
> me illustrate. We have all had the experience of driving without paying
> attention or of staring off into space. If attention were required for

> consciousness, it seems that we would not be phenomenally aware *in any
> sense* of the outside world. But this does not seem to me to be the case
> when I reflect on my own experience. In these cases, I would want to say
> that I was limitedly visually conscious (of course I could not describe
> the detail of my visual environment) even though my attention is directed
> elsewhere. If for example, the whole world were to disappear in one of
> these 'reflective' moments I would be more inclined to say that I was
> conscious of the world around me and the change directed my attention to it
> then that I was not conscious of the world until it's 'disappearance'.

I like this proposal for two reasons. First, it allows attention and consciousness to be disassociated from one another. Up until our last class meeting, I never thought about the possibility that these two aspects could be decoupled. However in light of the class discussion I now find this notion to be very intriguing. It brings up the possibility that consciousness may be divided into two categories: attentionally focused and non-attentionally focused. Attentionally focused consciousness would be the consciousness of contents that are within the attentional spotlight of one's visual field. It is within this spotlight that detailed visual information would be perceived and made globally accessible to the rest of the processing in the brain. Non-attentionally focused consciousness would be the consciousness of contents outside the attentional spotlight. Here there would be a considerable lack of detailed visual information, yet the information would still be globally available. Such a distinction may also tie in to discussion of conscious vs. unconscious processing in that it may be that certain processes that are usually considered unconscious may in fact be conscious but do not occur within the spotlight of attentional focus.

The second reason why I like Rachel's proposal is that it still allows that internal representations of visual stimuli are being produced by the visual system. The central idea of the enactive view - that no internal representations are necessary for consciousness- I find difficult to swallow. If a percept is conscious in virtue of the fact that it is externally available, how does this account for imagination and internal visual imagery? One could close their eyes and recreate in the mind's eye the experience of a perceptual stimulus. Such a percept would no longer be externally available, yet still will be robustly conscious. The fact that such contentful imagery is conscious seems to be a counterargument to the idea that no internal representation is necessary for conscious contents. If in the case of imagination the brain is not creating an internal representation, then just what is going on? I suppose that one could argue that the case in which one views a stimulus with eyes open is different from the case in which one imagines the stimulus; perhaps the various correlations in brain activity that accompany visual perception are just registrations or acknowledgements of the presence and nature of visual stimuli and not representations per se. However there is much empirical evidence that seems to indicate that the same systems that are involved in visual perception are also involved in visual imagery. To suggest that these systems are representing in one case and not representing in the other flies in the face of providing a parsimonious explanation. However Rachel's idea is parsimonious in that internal representation would be occurring in both cases, where the detail of representation is a function of the attention allocated to the features of the stimulus.

Logan T.

From landsurveyor@hotmail.com Sun Oct 17 16:57:45 1999

Return-Path: <landsurveyor@hotmail.com>

Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA09030
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 16:57:44 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JH8YB62A0G9D7MM2@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
17 Oct 1999 16:58:11 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JH8YB5KZ9S9EEZLX@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 17 Oct 1999 16:58:11 -0700 (MST)

Received: from hotmail.com (f323.hotmail.com [207.82.250.248])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with SMTP id QAA09021 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 16:57:40 -0700

Received: (qmail 7775 invoked by uid 0); Sun, 17 Oct 1999 23:57:37 +0000

Received: from 128.196.51.171 by www.hotmail.com with HTTP; Sun,
17 Oct 1999 16:57:37 -0700 (PDT)

Date: Sun, 17 Oct 1999 16:57:37 -0700 (PDT)

From: Matt Herbert <landsurveyor@hotmail.com>

Subject: attention, memory and consciousness

To: scicon@paradox.soc-sci.arizona.edu

Message-id: <19991017235737.7774.qmail@hotmail.com>

MIME-version: 1.0

Content-type: text/plain; format=flowed; CHARSET=US-ASCII

X-Originating-IP: [128.196.51.171]

Status: RO

All,

I had some residual questions/worries from the discussion of attention and
consciousness last week, so here goes.

Hardcastle evades the problem of change blindness across saccades by
suggesting that incoming visual data displaces the old data from iconic
memory so quickly that the subject simply cannot remember them to report
them. What looked like a gap in consciousness turns out to be a failure of
memory. I'm wondering what this view, if correct, entails for the
phenomenal present. If sensory experiences are so fleeting as to be pushed
out of iconic memory in as little as 250msec, it would seem that all our
so-called perceptual reports are actually mnemonic reports. (A question for
those in the know: Is iconic memory the same mechanism as the "buffer" for
a given modality? It appears by all descriptions that it is, but I cannot
track down a definitive answer.) Implications for a theory of consciousness
are potentially significant. If our visual experience is so transient, and
it is always the fate of our visual data to be stored in iconic memory,
perhaps consciousness will be nothing more than the ephemeral contents of
iconic memory. This doesn't sound very likely to me, but it is a possible
result of Hardcastle's line of thought.

Hardcastle's alternative explanation of Right Parietal Extinction and
Balint's syndrome patients' symptoms suggest that she herself would not
favor a conception of consciousness on which its contents are so highly
processed as to make it into iconic memory. The consensus explanation of
the Extinction patients' symptoms is that the apparent visual deficits are
due to attentional deficits. Hardcastle suggests, however, that even the
apparent attentional deficits are not enough to show that there are no
objects of perceptual experience. There might be something being
experienced that is not sufficiently processed (integrated, conceptualized)
to be reported. I don't have Hardcastle's text available at the moment, but
I believe she suggests that bare consciousness provides the inchoate
elements of experienced objects (unintegrated features), but attentional
processes provide the objects themselves. I'm not sure this is a plausible
defense. If there is a proto-experience that underlies full-blown

experience, it seems to be not quite what is meant by "the contents of consciousness." To speak only of my own phenomenology, I do not experience unintegrated features, but integrated objects. I believe as firmly as Hardcastle that there are data present in my phenomenal field that are, in a sense, prior to the neatly-bounded objects of my experience, but I *infer* their presence; I do not experience it. In short, I think it's fair to say that something like James' blomming buzzing confusion underlies one's phenomenal experience, but such proto-experience is *less* properly called experience than its more processed relatives.

Matt Herbert Nim eat Nim eat.
 Drink eat me Nim.
 Me gum me gum.
 You me banana me banana you.
 --Nim the signing chimp

Get Your Private, Free Email at <http://www.hotmail.com>

From lnielsen@azstarnet.com Sun Oct 17 17:49:41 1999
Return-Path: <lnielsen@azstarnet.com>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA09225
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 17:49:40 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JH904JZ3U89D7MFK@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
17 Oct 1999 17:50:07 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JH904JQ7W09EF1V9@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 17 Oct 1999 17:50:07 -0700 (MST)
Received: from cepheus.azstarnet.com (cepheus.azstarnet.com [169.197.56.195])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA09221 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 17:49:35 -0700
Received: from zippo (dialup001ip091.tus.azstarnet.com [169.197.12.91])
by cepheus.azstarnet.com (8.9.3+blt.Beta0/8.9.3) with SMTP id RAA09523 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 17:49:59 -0700 (MST)
Date: Sun, 17 Oct 1999 17:55:52 -0700
From: Lis Nielsen <lnielsen@azstarnet.com>
Subject: RE: iconic memory and buffers
In-reply-to: <19991017235737.7774.qmail@hotmail.com>
To: scicon@paradox.soc-sci.arizona.edu
Reply-to: lnielsen@u.arizona.edu
Message-id: <NCBBJCIPCKAPOGDI CDGPOEEGCDAA.lnielsen@azstarnet.com>
MIME-version: 1.0
X-MIMEOLE: Produced By Microsoft MimeOLE V5.00.2314.1300
X-Mailer: Microsoft Outlook IMO, Build 9.0.2416 (9.0.2910.0)
Content-type: text/plain; charset=iso-8859-1
Content-transfer-encoding: 7bit
Importance: Normal
X-Priority: 3 (Normal)
X-MSMail-priority: Normal
X-Sent-via: StarNet <http://www.azstarnet.com/>
Status: RO

Matt asked:

Is iconic memory the same mechanism as the "buffer" for a given modality? It appears by all descriptions that it is, but

I cannot track down a definitive answer.

Iconic memory is the memory of the stimulus that persists in the sensory system for a very brief time after stimulus presentation. Visual afterimages are an example of iconic memory. In the auditory modality, one speaks of echoic memory. These 'memories' require no rehearsal in order to hang around. On the other hand, buffer storage refers to information in working memory that is rehearsed and available for ongoing processing. The phonological loop (in which you repeat the phone number you want to remember) or the visuo-spatial scratch pad (in which you store the memory of some visual image in order to access it again) are examples of types of buffers related to different modalities in Baddeley's working memory system. I don't have the Hardcastle paper with me to see if she uses the term buffer to refer to iconic memory, but the distinctions above are those that I am familiar with from Cognitive Psychology.

Lis

From chalmers@paradox.soc-sci.arizona.edu Sun Oct 17 21:38:33 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA09531
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 21:38:33 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JH984BSBC09D6WC0@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
17 Oct 1999 21:39:00 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JH984BDLVK9EF2BD@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 17 Oct 1999 21:38:59 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id VAA09524 for scicon; Sun, 17 Oct 1999 21:38:23 -0700
Date: Sun, 17 Oct 1999 21:38:23 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: Re: attention and consciousness
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199910180438.VAA09524@paradox.soc-sci.arizona.edu>
Status: R

Quick notes about attention and consciousness:

(1) Logan suggests that phenomena of imagination and imagery provide evidence against the "enactive" claim that internal representation is not required for consciousness. In reply, an enactivist could say: internal representation is **compatible** with consciousness, and may even be required for some **kinds** of consciousness (perhaps for imagery, imagination, attention, etc). The claim is just that it isn't necessary -- so there are some conscious states that can be present without internal representation. When the claim is viewed this way, these phenomena may no longer be evidence against it.

(2) Juraj asks about the difference between A-consciousness and attention. He notes that A-consciousness is dispositional but consciousness doesn't seem dispositional. One response is to say that P-consciousness certainly isn't dispositional, and maybe Juraj is just importing intuitions about P-consciousness here. A second would be to say that A-consciousness is **defined** in terms of accessibility, so it is dispositional by definition. A third response would be to say maybe Juraj is implicitly understanding A-consciousness to be defined

differently, so that access rather than just accessibility is required. That would be OK, but it would be a terminological point; A-consciousness is itself a term of art and we can define it how we like. I like to stress the definition in terms of accessibility, since this correlates best with P-consciousness. But if one is less concerned with this correlation, there is certainly some valid notion of consciousness -- call it A'-consciousness -- that requires not just accessibility but access. And it may well be that A'-consciousness is strongly connected to attention.

(3) Matt draws some interesting consequences from Hardcastle's point that phenomenal iconic memory is really short and that this can distort the accuracy of phenomenal reports. He says this might suggest that phenomenal perceptual reports are really mnemonic reports; he also suggests that maybe then consciousness is no more than iconic memory. These are interesting claims but I'm not quite sure why they follow.

Re the first claim, presumably there roughly two sorts of reports (of visual experiences, say): those made while the experience is still present and those made after the experience is absent. Presumably we'd say the latter are in some sense mnemonic and the former are in some sense perceptual. It's not clear to me how the shortness of memory will itself make reports in the first class mnemonic; on the face of it, this point would have more impact on those in the second class (those we already take to be mnemonic). If the experience is still present, the shortness of memory will presumably be fairly irrelevant, and the reports of the experience will be broadly perceptual. Of course there is processing time involved, which one might argue will introduce a mnemonic element, but it's not clear how the shortness of memory will make a difference to this point.

Re the second claim, maybe one could argue that our reports of consciousness are really reflecting our iconic memory (somewhat as in the last point above). But presumably Hardcastle would use this to suggest that the reports don't fully reflect consciousness itself, as there is a difference between consciousness and memories thereof. It's not clear to me how one gets from this observation to the claim that all there is to consciousness is iconic memory. But maybe Matt can elaborate on the reasoning here.

Re Matt's point about pre-integrated experience: it's true that in familiar cases, where there is integration of experience, we count the integrated objects as part of experience, not the pre-integrated stuff. But that's compatible with saying that in fringe or fleeting cases where the experience never gets integrated at all, the pre-integrated stuff is still part of experience. It's as if experience can be processed and developed to various levels. The sort that we are most familiar with (from attention etc) is the integrated sort, but the other sort might also exist, even if for obvious reasons it is natural that it is not so familiar.

--Dave.

From jmartine@U.Arizona.EDU Sun Oct 17 22:12:15 1999

Return-Path: <jmartine@U.Arizona.EDU>

Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id WAA09609

for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 17 Oct 1999 22:12:14 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)

id <01JH99A46TI89D7R8W@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@ARIZONA.EDU); Sun,
17 Oct 1999 22:12:42 MST

Received: from orion.U.Arizona.EDU by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
with ESMTP id <01JH99A3SFGG9EF3ED@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Sun, 17 Oct 1999 22:12:41 -0700 (MST)

Received: from localhost (jmartine@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id WAA05450; Sun,
17 Oct 1999 22:12:40 -0700 (MST)

Date: Sun, 17 Oct 1999 22:12:39 -0700 (MST)

From: Joel A Martinez <jmartine@U.Arizona.EDU>

Subject: Re: Consciousness and attention, etc.

In-reply-to: <Pine.HPX.4.10.9910141059170.630-100000@pavo.U.Arizona.EDU>

To: Lonnie A Nelson <lan@U.Arizona.EDU>

Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu

Message-id: <Pine.HPX.4.10.9910172114270.12195-100000@orion.U.Arizona.EDU>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: R

I think Rachael's suggestion sounds interesting. Ultimately, though, I think to defend this view one would have to answer some version of Dave's question. (Namely, how coarse-grained is consciousness outside attention?) This is because this view walks a fine line between the Dennett/Mack-Rock view and the Wolf/Hardcastle views.

A comment Logan made drew attention to an aspect of Rachael's suggestion that I think is interesting and potentially problematic. Logan has a favorable attitude towards this view b/c it separates consciousness and attention. So, non-attentionally focused consciousness would be very coarse grained but still globally available, in some sense. But, I think its worth asking in what sense non-attentionally focused consciousness is available. The change blindness studies seem to show that the information is not available. Subjects cannot use it to help guide action, they do not have any memory of something passing in front of thier visual field, etc... So, in what sense is the information available? If we are going to call the non-attentionally focused consciousness conscious should'nt we be able to say something here?

I think this is just a version of Dave's question. (To what extent do we notice changes in the periphery of our visual field?) B/c I am really just trying to put some pressure on the idea that we can be conscious of what we are not attending to.

As Rachael points out, the cocktail party phenomena is relevant here. She comments that conscious processes must be going on if one is going to notice, through all the background noise, someone mention her name. This, along with Lonnie's idea of "attentional sharpness" (where she briefly mentions how one mode, say vision, may not be so "sharp" in some individuals), started me thinking about the relevance of other modes of percepton here. We have different senses at work constantly. It seems likely to me that they influence one another (e.g. hearing affects vision, in some way), and each acts in its own way to help produce a full conscious experience. While this does not help with the question of how coarse-grained the periphery of our visual field is, it may help to explain how our experience can seem unified. That is, there is a lot more than visual input that is relevant when discussing the "smooth-ness" and completeness of conscious experience. I guess what I am wondering is whether information received through modes other than our visual mode of perception is relevant in considering the "smooth-ness" of *visual consciousness*. I am not quite sure how this might work yet. Just a thought. (This is a bit off topic, but I hope its not too unrelated.)

-Joel M.

Joel A. Martinez
Social Science Bldg. Rm. 213
Dept. of Philosophy
University of Arizona
P.O. Box 210027
Tucson, AZ 85721-0027
USA

Office #138 Phone 520-621-7098

From rachaelp@U.Arizona.EDU Tue Oct 19 10:24:12 1999
Return-Path: <rachaelp@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id KAA13384
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 19 Oct 1999 10:24:12 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.2-31 #39830)
id <01JHBD5088009GVF6Q@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Tue,
19 Oct 1999 10:24:41 MST
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU)
by Telcom.Arizona.EDU (PMDf V5.2-31 #39830)
with ESMTP id <01JHBD4ZSMDC9JDCYA@Telcom.Arizona.EDU> for
chalmers@Ariz0na.EDU; Tue, 19 Oct 1999 10:24:41 -0700 (MST)
Received: from localhost (rachaelp@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id KAA08077; Tue,
19 Oct 1999 10:24:40 -0700 (MST)
Date: Tue, 19 Oct 1999 10:24:40 -0700 (MST)
From: Rachael J Parkinson <rachaelp@U.Arizona.EDU>
Subject: Re: Consciousness and attention, etc.
In-reply-to: <Pine.HPX.4.10.9910172114270.12195-100000@orion.U.Arizona.EDU>
To: Joel A Martinez <jmartine@U.Arizona.EDU>
Cc: Lonnie A Nelson <lan@U.Arizona.EDU>, David Chalmers <chalmers@Ariz0na.EDU>,
scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910190934530.10991-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Joel writes:

> I think Rachael's suggestion sounds interesting. Ultimately, though,
> I think to defend this view one would have to answer some version of
Dave's question. (Namely, how coarse-grained is consciousness outside
attention?) This is because this view walks a fine line between the
> Dennett/Mack-Rock view and the Wolf/Hardcastle views.

This question echoes Dave's concern when he asks just how coarse-grained
the consciousness outside attention is. I think my view is closer to that
of Wolfe and Hardcastle than to Dennett/Mack-Rock. On the 'grand
illusion' view there is no consciousness outside attention. I would

disagree with this. Dennett/Mack and Rock have a claim about internal representation that is similar to mine, that is, that it is limited in some way. But the difference is this: For them, internal representation is limited to what we are attending to. On my view, there is internal representation, albeit course grained (and therefore 'limited') outside of attention. So we are pointing to different things when we suggest that our internal representation is limited. This view differs from the Wolfe/Hardcastle view only in that it recognizes that consciousness outside of attention is 'fuzzy' and does not have the detail that consciousness within attention does.

Joel goes on to point out:

> A comment Logan made drew attention to an aspect of Rachael's
> suggestion that I think is interesting and potentially problematic. Logan
> has a favorable attitude towards this view b/c it separates consciousness
> and attention. So, non-attentionally focused consciousness would be very
> coarse grained but still globally available, in some sense. But, I think
> its worth asking in what sense non-attentionally focused consciousness is
> available. The change blindness studies seem to show that the information
> is not available. Subjects cannot use it to help guide action, they do
> not have any memory of something passing in front of thier visual field,
> etc... So, in what sense is the information available? If we are going to
> call the non-attentionally focused consciousness conscious should'nt we be
> able to say something here?
> I think this is just a version of Dave's question. (To what extent do we
> notice changes in the periphery of our visual field?) B/c I am really just
> trying to put some pressure on the idea that we can be conscious of what
> we are not attending to.

I guess I think that my view can accommodate the change blindness studies because those things that are being changed are pretty fine grained (take the lines in the road or the engine on the plane, for example). I do not think that we are aware of the engine or the lines in the road when we are not attending to them. Thus, the awareness that there is an engine is most likely not globally available. Consciousness outside attention is much more course grained than this.

Recall that in my original e-mail I suggested that consciousness outside of attention may be so course grained that the most that we can say of it is that we recognize it (our fuzzy phenomenal experience) only when it is missing. Of course, it would be nice to discover that consciousness outside attention is not quite so course grained as that; perhaps we are aware of color, shape, or particularly large or 'intrusive' aspects of our inattentional conscious field. ('intrusive' like our names at a cocktail party.)

It would be interesting to develop further change blindness studies to see if we are aware of particular course grained features of our unattended consciousness. Of course, I find the arguments that there may be limited or no memory outside of attention persuasive. Even if we were to find that there are no course grained features of consciousness that subjects recognize in change blindness studies, we would have to rule out the possibility that the subjects were conscious of but could not remember those course grained features.

I hope these points help to clarify my view.

-Rachael

From sawright@U.Arizona.EDU Sun Oct 24 20:08:51 1999
Return-Path: <sawright@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id UAA23514
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 24 Oct 1999 20:08:51 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHIX0RJH8G9GXFON@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
24 Oct 1999 20:09:29 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHIX0R0XLC9JFHI5@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 24 Oct 1999 20:09:28 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id UAA23510 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 24 Oct 1999 20:08:30 -0700
Received: from localhost (sawright@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id UAA17440 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 24 Oct 1999 20:09:08 -0700 (MST)
Date: Sun, 24 Oct 1999 20:09:08 -0700 (MST)
From: Sarah A Wright <sawright@U.Arizona.EDU>
Subject: Processing and Anatomy
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910241949580.7654-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

A note about the connection between mental processing and brain anatomy:

Baars is quick to point out that his model of the processing involved in consciousness does not necessarily closely correspond to our best present neurological pictures. This is not a requirement for his model since it is meant to capture the level of processing, not of neuroanatomy. So he is allowing, quite judiciously, that the organization of the brain into processing systems need not follow any sort of gross brain grouping. A process can be spread out all over the brain, or isolated into a small area, and this has no effect on the processing model.

This is in sharp contrast with Schacter, who uses current theories of neuroanatomy to motivate the distinction between the CAS and the Executive system. One is in the posterior parietal cortex and the other is in the prefrontal cortex; they can't be the same thing! Of course, making such an argument requires a reason to think that processing follows gross grouping, and we haven't seen a reason to accept that yet.

I have a question and a comment about this issue.

1) Does anyone know exactly how the evidence is developing in other areas besides consciousness? Is there reason to think that simple brain processes are quite diffuse?

Is the focus on finding processes for brain areas driven by a prediction that processes will be so localized, or is it merely an artifact of the limitations of current brain mapping technology?

2) If there is reason to believe that brain processes are diffuse, that seems to be at odds with the methodology and results that we have seen in the search for the NCC, which seemed to focus on finding a particular location, and which differed primarily in the size of the NCC and the location chosen.

The methodology in those experiments focused on causing the mental process

to be tested (e.g. conscious perception), then looking to a particular brain region for any signs of neural activity. In the experiments in which a single cell was monitored, the results from that one cell were taken to indicate the activity of a general area of cells; such a generalization would require a strong thesis that neural grouping follows mental processing.

So perhaps we should look back at the results about the NCC from the beginning of the class to see what has actually been demonstrated if we do not assume that all cells in a brain area are engaged in the same sorts of processing activities.

Any thoughts on this problem?

Sarah

From press@U.Arizona.EDU Sun Oct 24 22:13:12 1999
Return-Path: <press@U.Arizona.EDU>
Received: from Arizona.EDU (Hopey.Telcom.arizona.edu [128.196.128.234])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA23860
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 24 Oct 1999 22:13:11 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHJ1CXRGV49GXGE0@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
24 Oct 1999 22:13:49 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHJ1CXC9V49JF49S@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 24 Oct 1999 22:13:49 -0700 (MST)
Received: from f1n3.u.arizona.edu
(IDENT:press@f1n3.U.Arizona.EDU [128.196.137.103])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA23855 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 24 Oct 1999 22:13:01 -0700
Received: from localhost (press@localhost) by f1n3.u.arizona.edu (8.8.8/8.8.8)
with ESMTP id WAA31818 for <scicon@paradox.soc-sci.arizona.edu>; Sun,
24 Oct 1999 22:13:38 -0700
Date: Sun, 24 Oct 1999 22:13:38 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>
Subject: Verificationism
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9910242211230.44974-100000@f1n3.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

All,

First, a confession. During my formative years as a philosopher, I have been rather heavily exposed to both Dennett and verificationism, and there are soft spots in my heart for both of them. In general, I do not think this has lead to corresponding soft spots in my head, but full disclosure nevertheless seemed appropriate. With that in mind, a few thoughts about Dennett, verificationism, and consciousness.

Even if we agree that verificationism is flawed as a theory of meaning, the related epistemological lessons are worthwhile in themselves. I think Simon was right to point out that the issues Dennett and Kinsbourne raise only really show that certain theories of consciousness are underdetermined by the data. But isn't this a fairly serious problem in its own right? Several of the issues we've been discussing lately seem

to be in this situation. Is the color-phi phenomenon due to an Orwellian revision of memory or a Stalinesque manipulation of the perceptual data? Is inattentional blindness the result of a lack of conscious perception or near-instantaneous memory loss? Do blindsight patients have qualia, and if so, how faded and/or distorted are they? These questions all seem to be unanswerable because any actual or imaginable data appears to be fully and equally consistent with all the answers.

Now, a full-fledged verificationist would claim that this makes these questions meaningless. This seems wrong, because even if we cannot decide between them based on the available data, they certainly seem to be conceptually distinct possibilities. But my point would be that even if there is a fact of the matter here, if nothing we can ever observe will verify our acceptance of one or the other of these theories, what is the point of endorsing any theory about these phenomena at all? We can, of course, decide to take a stand on one side of these issues, and others could decide to take a stand on the opposite sides, but on what grounds should the two opposing camps argue? It seems that any evidence offered in support of one side could equally well be used by the opposition. If so, we have guaranteed the existence of a perpetual philosophical argument, which, though possibly good for individual philosophers, is bad for philosophy generally.

I'm not sure whether Dennett would agree with this analysis, but when he calls these sorts of issues, "differences that make no difference," I read this as meaning that even if there really is a difference, it wouldn't make us any smarter to believe the true theory than its alternative (we wouldn't be able to explain anything more with the theory than we could without it). Dennett seems to want to take things further by saying that such differences are not really differences at all, but either way, this seems to be bad news for a fair amount of what's been said about consciousness so far. If we really believe that there is a difference between Orwellian and Stalinesque revisions and so forth, it seems that we need to be looking at ways of teasing apart the verification conditions of the two if at all possible. The other possibility would be to admit that we are only preferring one theory to another for practical or conventional reasons, but if we really believe that there is an actual factual difference, this won't be satisfactory. Or, even less satisfactorily, we could give up, as I suggested Velmans was doing a couple of weeks ago.

I think there is probably more to be said on this subject, but the post has gotten rather long already, so I'll let it rest at this for now.

Verbosely,
Joel

From bradt@u.arizona.edu Sun Oct 24 23:18:56 1999

Return-Path: <bradt@u.arizona.edu>

Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA23985
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 24 Oct 1999 23:18:55 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHJ3NFFZDS9GXFAS@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
24 Oct 1999 23:19:33 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHJ3NF22U89JFJ26@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 24 Oct 1999 23:19:32 -0700 (MST)

Received: from hawk.prod.itd.earthlink.net
(hawk.prod.itd.earthlink.net [207.217.120.22]) by paradox.soc-sci.arizona.edu

(8.9.3/8.9.3) with ESMTTP id XAA23980 for <scicon@paradox.soc-sci.arizona.edu>;
Sun, 24 Oct 1999 23:18:40 -0700
Received: from mrector.shiva.com
(ip112.tucson6.az.pub-ip.psi.net [38.29.66.112])
by hawk.prod.itd.earthlink.net (8.9.3/8.9.3) with SMTP id XAA11670; Sun,
24 Oct 1999 23:19:12 -0700 (PDT)
Date: Sun, 24 Oct 1999 23:23:35 -0700
From: Brad Thompson <bradt@u.arizona.edu>
Subject: Comments on Baars
To: scicon@paradox.soc-sci.arizona.edu
Cc: Brad Thompson <bradt@u.arizona.edu>
Reply-to: bradt@u.arizona.edu
Message-id: <000001bf1eb1\$78c0a900\$70421d26@mrector.shiva.com>
MIME-version: 1.0
X-MIMEOLE: Produced By Microsoft MimeOLE V4.72.3110.3
X-Mailer: Microsoft Outlook 8.5, Build 4.71.2173.0
Content-type: text/plain; charset=iso-8859-1
Content-transfer-encoding: 7bit
Status: R

I found the Baars article from last week unsatisfying, so here are a few criticisms that come to mind (in addition to those we discussed in class).

1. "...we can create access to any part of the brain using consciousness"

Baars supports this bold claim with the example of biofeedback techniques in which subjects are able to "control" the behavior of even single neurons if given feedback regarding the activity of the neuron (such as a tone). But I don't see what this has to do with consciousness (other than the fact that subjects are conscious human subjects and that they happen to be conscious of the tones, etc.). And it is certainly misleading to say that this ability shows that we have "access" to the particular brain region under "control". If he means by "access" something like consciousness or global availability, it is surely false. And it isn't clear that consciousness is playing a critical role in these cases. I would think that simple behaviorist mechanisms could be at work here. Give a rat yummy food when it is generating high levels of alpha waves, and shock it when it is not, and I'll bet that it will gain "control" too. Maybe I'm wrong about this prediction, but it is still true that Baars doesn't do a very good job of motivating the view that some special property of consciousness is at work here. Rather, the "control" at issue seems extremely indirect and unconscious.

Likewise with his example of changes in our understanding of the meaning of the word "set" depending on the conscious hearing of another word, such as "tool". The changes in the meaning of "set" is due to the activation of a semantic representation of "tool". Now maybe semantic representations don't get activated unless the word "tool" is heard consciously. But we can't conclude from this as Baars does that "consciousness is required to integrate the meaning of each word pair." Rather, you have to *hear* (consciously) a word in order for its semantic representation to be activated, and the latter has an effect on the interpretation of subsequent ambiguous words. Consciousness is playing an incidental, not an essential role here.

The above criticisms apply as well to some of Baars' other examples, such as memory for visual images.

2. "The Universality of theatre models"

Baars seemed to be trying to bolster his theatre metaphor by pointing out how often people appeal to it. That it is a horrible way to argue for a

position. And it should be pointed out that his citing of Plato's allegory of the cave was way off the mark as an example of a theatre model of consciousness.

3. The function of consciousness

I found the following passage weird. "Consciousness creates access to unconscious problem-solving. The famous "incubation process" in mathematics involves a conscious question, unconscious work on the problem, and a conscious emergence of the solution." Baars takes this as another example of the function of consciousness being to present in the theatre something for the (unconscious) audience to do work on. But again, as in my criticisms above (1), why think that consciousness per se is doing the work here? And why would this support the theatre metaphor anyway? Maybe the question posed must have some threshold of activation in order to trigger some further process of problem solving--but this can all be explained without positing some place (physically or merely functionally specified) where it all comes together.

One final problem. Baars says in his conclusion that consciousness *creates* "access to many independent knowledge sources in the brain, most of them quite unconscious". But in the next paragraph, he says that global access may be a necessary condition for consciousness. This would support the view that consciousness *is* a special kind of access, not that it *creates* access. It looks like Baars is confused between these two different views. I think that the rest of the paper supports the "create access" reading, since that is more in line with his view that consciousness has the function of making information available to unconscious processes.

From lan@U.Arizona.EDU Mon Oct 25 12:00:37 1999

Return-Path: <lan@U.Arizona.EDU>

Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id MAA25329
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 25 Oct 1999 12:00:37 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHJU9SQ0K09GX6HP@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Mon,
25 Oct 1999 12:01:16 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHJU9SCBRK9JF60P@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 25 Oct 1999 12:01:15 -0700 (MST)

Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id MAA25324 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 25 Oct 1999 12:00:19 -0700

Received: from localhost (lan@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id MAA18099; Mon,
25 Oct 1999 12:00:57 -0700 (MST)

Date: Mon, 25 Oct 1999 12:00:57 -0700 (MST)

From: Lonnie A Nelson <lan@U.Arizona.EDU>

Subject: Re: Verificationism

In-reply-to: <Pine.A41.4.10.9910242211230.44974-100000@f1n3.u.arizona.edu>

To: Joel K Press <press@U.Arizona.EDU>

Cc: scicon@paradox.soc-sci.arizona.edu

Message-id: <Pine.HPX.4.10.9910251135510.862-100000@pavo.U.Arizona.EDU>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: RO

On Sun, 24 Oct 1999, Joel K Press wrote:

Joel summarizes the verification and underdetermination of theory by data below. It seems to me that the problems of Orwellian revision and Stalinesque Manipulation of perceptual data each suggest a "possible world" which are very different worlds. It would seem based on the existence of these possible worlds that it is also likely that if these two issues are different in these worlds, they would imply further differences in the worlds. These further differences would likely be 1) either resultant of the nature of a conscious system that operates this way (Orwellian or stalinesque) also showing the same sort of processing in other modalities and time scales that is at least more consistent with one story than the other. So the question to ask here is: IF this variable is operating according to story X, what else in this world should be different and in what way?

OR 2) determined by two or more possible structures in disparate function of the instantiating system in each of these worlds. So the question here is: IF Story x is true, we should see the memory centers behave in way x, if story y is true, we should see perceptual systems behave in way y; which way do these things behave?

This view of the problem admits the impossibility of choosing between the two possible worlds based on the single story/case of the color Phi instance with the data already available, but also hypothesizes that there would be other findings that should be more consistent with one overall depiction of how consciousness operates under various conditions than the other story. So at the end of the day we are not choosing an entire story of consciousness based on this one example, but we are choosing the most consistent story of this example based on that story's fit with our overall story of consciousness.

Thus one avoids giving up, by looking for other possible supporting evidence that is more consistent with one overall story than the other but not directly applicable to the smaller scale question that one desires a direct explanation for. Does this make sense to anyone but me?

At least equally verbosely,
--Lonnie

> Even if we agree that verificationism is flawed as a theory of
> meaning, the related epistemological lessons are worthwhile in themselves.
> I think Simon was right to point out that the issues Dennett and
> Kinsbourne raise only really show that certain theories of consciousness
> are underdetermined by the data. But isn't this a fairly serious problem
> in its own right? Several of the issues we've been discussing lately seem
> to be in this situation. Is the color-phi phenomenon due to an Orwellian
> revision of memory or a Stalinesque manipulation of the perceptual data?
> Is inattentional blindness the result of a lack of conscious perception or
> near-instantaneous memory loss? Do blindsight patients have qualia, and
> if so, how faded and/or distorted are they? These questions all seem to
> be unanswerable because any actual or imaginable data appears to be fully
> and equally consistent with all the answers.

>
> Now, a full-fledged verificationist would claim that this makes these
> questions meaningless. This seems wrong, because even if we cannot decide
> between them based on the available data, they certainly seem to be
> conceptually distinct possibilities. But my point would be that even if
> there is a fact of the matter here, if nothing we can ever observe will
> verify our acceptance of one or the other of these theories, what is the
> point of endorsing any theory about these phenomena at all? We can, of
> course, decide to take a stand on one side of these issues, and others
> could decide to take a stand on the opposite sides, but on what grounds
> should the two opposing camps argue? It seems that any evidence offered
> in support of one side could equally well be used by the opposition. If

> so, we have guaranteed the existence of a perpetual philosophical
> argument, which, though possibly good for individual philosophers, is bad
> for philosophy generally.
>
> I'm not sure whether Dennett would agree with this analysis, but when he
> calls these sorts of issues, "differences that make no difference," I read
> this as meaning that even if there really is a difference, it wouldn't
> make us any smarter to believe the true theory than its alternative (we
> wouldn't be able to explain anything more with the theory than we could
> without it). Dennett seems to want to take things further by saying that
> such differences are not really differences at all, but either way, this
> seems to be bad news for a fair amount of what's been said about
> consciousness so far. If we really believe that there is a difference
> between Orwellian and Stanlinesque revisions and so forth, it seems that
> we need to be looking at ways of teasing apart the verification conditions
> of the two if at all possible. The other possibility would be to admit
> that we are only preferring one theory to another for practical or
> conventional reasons, but if we really believe that there is an actual
> factual difference, this won't be satisfactory. Or, even less
> satisfactorily, we could give up, as I suggested Velmans was doing a
> couple of weeks ago.
>
> I think there is probably more to be said on this subject, but the post
> has gotten rather long already, so I'll let it rest at this for now.
>
> Verbosely,
> Joel
>
>
>

It is a common fate of all knowledge to begin as heresy and end as
orthodoxy.

-Thomas Huxley

Lonnie A Nelson
Department of Psychology
Human Energy Systems Laboratory
University of Arizona
lan@u.arizona.edu

From anhabib@U.Arizona.EDU Mon Oct 25 13:09:46 1999
Return-Path: <anhabib@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id NAA25542
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 25 Oct 1999 13:09:45 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHJW0J44I09GXJH1@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Mon,
25 Oct 1999 13:10:24 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHJW0IP7609JD49C@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 25 Oct 1999 13:10:24 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id NAA25538 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 25 Oct 1999 13:09:30 -0700
Received: from localhost (anhabib@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id NAA27225 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 25 Oct 1999 13:10:09 -0700 (MST)

Date: Mon, 25 Oct 1999 13:10:09 -0700 (MST)
From: Allen N Habib <anhabib@U.Arizona.EDU>
Subject: Dennett et. al. (fwd)
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910251308231.19303-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R0

----- Forwarded message -----

Date: Fri, 22 Oct 1999 17:14:18 -0700 (MST)
From: Allen N Habib <anhabib@U.Arizona.EDU>
To: scicon@paradox.soc-sci.edu
Subject: Dennett et. al.
(sorry Dave, I initially sent this to the incorrect address above)

Hey all,

Last discussion left me hankering for a clarification of Dennett and Kinsbourne's argument. I'll try to give one, and use it to examine what I think are some problems with their thesis:

0. Cartesian theories of consciousness can offer two explanations of the abberant phenomena that D&K adduce

1. Both the Orwellian and Stalinesque theories adequately explain both the first person and third person data of the various phenomena that D&K adduce (colour phi, sub cutaneous rabbit, etc).

2. But the two theories posit different (non-isomorphic) mechanisms for the phenomena.

3. No empirical test, actual or possible, could tell the difference between these two hypotheses.

4. Any two theories that posit such an untestable difference are suspect.

Conclusion: Cartesian theories of consciousness are suspect

I can see two major problems with this line. The first one is with premis three. Arguably it is the case that current science has not found a way to tell which of these two stories is the correct one, but what proof do D&K proffer for their modal claim, that no evidence could ever do this job? Precious little. They do provide an example of the way a Stalinesque theorist might defend against a piece of data that imperils their theory, but this fals as a demonstration of the plausibility of the infinite room available for such moves.

Consider, say D&K, what would happen if the subject were asked to signal the moment that they become conscious of the first flash in the colour phi experiment. Further, imagine if this signal came in advance of the appearance of the second flash. It might seem that this is a decisive piece of evidence in favour of the Orwellian theory, since, regardless of the memory trace or subsequent verbal report of the subject, we have proof that the sunject was conscious of the first flash as a stationary red point, since such consciousness was reported before the advent of the second flash. Not so, say D&K, because the Stalinesque theorist might reply by saying that the subject in such a case signaled of the event before she was actually conscious of it. How? Well, the conscious mind of

the subject might have 'leaked' the instruction to push the button to the 'editing room' (the locus posited by Stalinists where the editing of conscious experience gets done), and the editing room pre-consciously initiated the button push.

But, as we can plainly see, this greatly sophisticates the Stalinesque theory. It requires that the editing room be able to receive instructions from the consciousness center, that the editing room have executive control over the motor cortex, etc. Also, if the experiment is done in such a way that the subjects are instructed not to push the button unless they have a conscious experience (and the results are the same), then the Stalinist must posit that the editing room acts upon the 'leaked' command in spite of the conscious determination not to on the part of the subject. Spooky, and more than a little ad hoc. This is not to say that the Stalinist couldn't say such things, but rather to point out that only rarely does one definitive test separate two competing theories. Normally what happens is that over a series of experiments a preponderance of evidence builds up in favour of one or another hypothesis. And while there is some room for a theorist to make amendments to her theory in the face of recalcitrant data, there isn't an infinite amount, and eventually the theorist is forced to either give up the theory or to make it completely ad hoc. D&K's contention that no evidence could ever surface that would definitively decide between these two theories seems false on its face.

Also, as was discussed in class, there is a problem for the fourth premis. This is the verificationist premis, and I have attempted here to provide the least contentious version of it that I could muster. As it stands, I don't think it is too objectionable, but as a result of its weakness, it doesn't grant D&K the conceptual purchase that they might like on cartesian materialism. If we grant that these two theories couldn't ever be decided between on the basis of empirical evidence, then we might well have a reason to question them, or, more specifically, to question the theoretical entities that underly the conflict. In this case, it would be the cartesian theatre (finish line of consciousness) that D&K would have us question. But D&K's answer, i.e. throw out the idea of a cartesian theatre altogether, seems a bit hasty. If we could simply amend the theatre idea in such a way as to solve D&Ks problems, but leave the theatre intact, then we wouldn't have to by their new product.

And we might well accomplish this (as we said in class) by making the temporal borders of the cartesian theatre 'fuzzy'. If we say that there is a brief span of time in the course of a percept's journey from unconsciousness to consciousness when there is no determinate fact of the matter whether that percept is conscious or not, then we can solve all of D&Ks temporality problems without throwing out the cartesian baby with the bath. Also, such temporal fuzziness is of a piece with the spatial (regional?) fuzziness of consciousness evidenced by peripheral and unattended conscious percepts. If consciousness' spatial borders are blurred, why not its temporal ones, as well? Any thoughts?

(Sorry about the length)

Al

From franzen@U.Arizona.EDU Mon Oct 25 08:40:43 1999
Return-Path: <franzen@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id IAA24740
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 25 Oct 1999 08:40:43 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHJN9YQKGW9GW97J@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Mon,
25 Oct 1999 08:41:21 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHJN9YD1TS9JFPDE@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 25 Oct 1999 08:41:21 -0700 (MST)
Received: from f1n7.u.arizona.edu
(IDENT:franzen@f1n7.U.Arizona.EDU [128.196.137.107])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id IAA24735 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 25 Oct 1999 08:40:29 -0700
Received: from localhost (franzen@localhost)
by f1n7.u.arizona.edu (8.8.8/8.8.8) with ESMTP id IAA36970 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 25 Oct 1999 08:41:08 -0700
Date: Mon, 25 Oct 1999 08:41:08 -0700 (MST)
From: Peter L Franzen <franzen@U.Arizona.EDU>
Subject: HOT
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9910241729290.38300-100000@f1n7.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Rosenthal writes, "We can reasonable assume that many of these children's mental states are conscious. But on the HOT hypothesis those states can be conscious only if the children have HOTS about them." I'm a bit confused about what it means to have an experience of something and to _not_ have a HOT about it, or is Rosenthal just saying that the young kids actually have HOTS that are not verbally expressed. What is the difference between having a conscious experience and having a HOT about it? Can one happen independently of the other? If three year old children do not have HOTS, to what extent, then, are they conscious of thoughts, sensations/perceptions, and the like?

Additionally, Rosenthal states that, "even adult HOTS are expressing their nonconscious HOTS in words. Indeed, although almost all verbally expressed thoughts are conscious, HOTS are an important exception. ... Not being conscious does not prevent HOTS from being verbally expressed." But, once a thought is expressed, isn't it then conscious? I'm not sure what else it would be.

As to Rosenthal's point about how false HOTS might be used to explain repressed beliefs and desires, I do not follow his argument. "Erroneous HOTS may well also figure in cases of so-called self-deception; there one's HOTS would misrepresent not what one desires but what one believes." This places too much omnipotence of HOTS. Sometimes, we just aren't good guessers (as research on human statistical inferences shows), perceivers, knowers of knowledge, etc.. Humans make a lot of mistakes, rely on heuristics, etc.. Thus, its not simply an erroneous HOT.

From landsurveyor@hotmail.com Sun Oct 31 14:11:30 1999
Return-Path: <landsurveyor@hotmail.com>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id OAA06767

for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 14:11:30 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHSCLE0DS9GZAA2@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
31 Oct 1999 14:12:17 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHSCLAYT609JHY6L@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 31 Oct 1999 14:12:16 -0700 (MST)
Received: from hotmail.com (f297.hotmail.com [207.82.251.189])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with SMTP id 0AA06762 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 14:11:23 -0700
Received: (qmail 11137 invoked by uid 0); Sun, 31 Oct 1999 21:11:41 +0000
Received: from 128.196.51.170 by www.hotmail.com with HTTP; Sun,
31 Oct 1999 13:11:41 -0800 (PST)
Date: Sun, 31 Oct 1999 13:11:41 -0800 (PST)
From: Matt Herbert <landsurveyor@hotmail.com>
Subject: HOTs and Consciousness
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <19991031211141.11136.qmail@hotmail.com>
MIME-version: 1.0
Content-type: text/plain; format=flowed; CHARSET=US-ASCII
X-Originating-IP: [128.196.51.170]
Status: RO

All--

A few notes about HOTs and consciousness, especially as treated by
Rosenthal. First, Peter's concern about children's consciousness. Peter
writes:

Rosenthal writes, "We can reasonable assume that many of these children's
mental states are conscious. But on the HOT hypothesis those states can
be conscious only if the children have HOTs about them." I'm a bit
confused about what it means to have an experience of something and to
not have a HOT about it, or is Rosenthal just saying that the young kids
actually have HOTs that are not verbally expressed.

Yes, Rosenthal (almost) says that children have the HOTs about experience
required for consciousness. A few lines after the passage cited by Peter,
Rosenthal says that the essential difference is between thought and report.
Even adults have thoughts that they cannot formulate as reports, so
presumably, children do too. Some such thoughts may be HOTs-- "So it may
well be that these children [three and younger] have the relevant HOTs even
though they do not express them verbally."

This line of thought has the interesting consequence of committing one to
nativism about mentalese if one is to maintain that infants are conscious.
If there's conscious experience, after all, it is so because it bears the
imprint of a thought, and a thought implicates at least one concept.
Regarding developing children who have not yet gained a large conceptual
repertoire (expressed in a natural language), nativism might be invoked to
explain the apparently rich conscious experience that children enjoy. It
would not be *required* to explain consciousness, but it does seem like the
natural direction to go if one is committed to nativism for infants.

On a separate note, Rosenthal poses one argument for the HOT theory as a
disjunctive syllogism: We have conscious content either by virtue of
sensing it or thinking about it. It cannot be the former; therefore, it is
the latter. This conclusion depends, of course, on the original
disjunction's being exhaustive and otherwise accurate.
For my part, I do not believe it to be either. I think Austin's argument
that we do not sense sense data (else a regress follows) is essentially

correct. We sense external objects that are internally represented. The internal representations, as far as I can see, do not require a further act of sensing to confer phenomenality on them. Perhaps Rosenthal would take this as evidence in his favor, since it is an argument against our sensing conscious content. However, I take it as an argument against his beginning supposition, since it makes one disjunct out to be not merely wrong, but implausible. To declare cognition as the gatekeeper of consciousness because it excels a plausible contender would be decent argumentation. However, if the competition is downright implausible and not clearly exhaustive (with cognition) of the field, a victory by default seems not to be decisive. (I must confess ignorance of Rosenthal's arguments for HOT cited in note 11.)

Matt Herbert Nim eat Nim eat.
 Drink eat me Nim.
 Me gum me gum.
 You me banana me banana you.
 --Nim the signing chimp

Get Your Private, Free Email at <http://www.hotmail.com>

From landsurveyor@hotmail.com Sun Oct 31 14:30:13 1999
Return-Path: <landsurveyor@hotmail.com>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id OAA06783
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 14:30:13 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHSD8J6DC09GZ0V7@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
31 Oct 1999 14:31:00 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHSD8INK289JHWQX@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 31 Oct 1999 14:30:59 -0700 (MST)
Received: from hotmail.com (f39.hotmail.com [207.82.250.50])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with SMTP id OAA06776 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 14:29:56 -0700
Received: (qmail 11789 invoked by uid 0); Sun, 31 Oct 1999 21:30:14 +0000
Received: from 128.196.51.170 by www.hotmail.com with HTTP; Sun,
31 Oct 1999 13:30:14 -0800 (PST)
Date: Sun, 31 Oct 1999 13:30:14 -0800 (PST)
From: Matt Herbert <landsurveyor@hotmail.com>
Subject: P.S.
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <19991031213014.11788.qmail@hotmail.com>
MIME-version: 1.0
Content-type: text/plain; format=flowed; CHARSET=US-ASCII
X-Originating-IP: [128.196.51.170]
Status: RO

P.S.--

It just occurred to me that I really didn't know what I meant in saying that a disjunction ought to be exhaustive and "otherwise correct." If it's exhaustive, the DS argument should go thru. So I guess what I mean about Rosenthal's disjunction is that it is simply not exhaustive. On my (Austinian) understanding, sensing and thinking are not jointly exhaustive, because I think there is a third option: *having* conscious experience. This will appear question-begging, since we are trying to give the criteria for *having* conscious experience, but suffice it to say for the moment that

I believe it can be had independently of second-order thoughts or sensations. To put it another way, I do not think that second-order mentation is required for conscious mentation. Hope that helps.

Matt

Get Your Private, Free Email at <http://www.hotmail.com>

From hvorecky@U.Arizona.EDU Sun Oct 31 15:17:52 1999
Return-Path: <hvorecky@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id PAA06819
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 15:17:51 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.2-31 #39830)
id <01JHSEWMDII89GZOPF@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
31 Oct 1999 15:18:40 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDf V5.2-31 #39830) with ESMTP id <01JHSEWM200W9JHHCU@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 31 Oct 1999 15:18:39 -0700 (MST)
Received: from f1n2.u.arizona.edu
(IDENT: hvorecky@f1n2.U.Arizona.EDU [128.196.137.102])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id PAA06814 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 15:17:43 -0700
Received: from localhost (hvorecky@localhost)
by f1n2.u.arizona.edu (8.8.8/8.8.8) with ESMTP id PAA42494 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 15:18:32 -0700
Date: Sun, 31 Oct 1999 15:18:32 -0700 (MST)
From: Juraj Hvorecky <hvorecky@U.Arizona.EDU>
Subject: Block and HOT
To: Science of Consciousness <scicon@paradox.soc-sci.arizona.edu>
Message-id: <Pine.A41.4.10.9910311455020.42206-100000@f1n2.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

The outcome of the last discussion has led me back to the talk Block was giving to the Consciousness Studies meeting couple weeks ago. For those who haven't been there, his message was somehow close to what we have discussed, but he said there is a way how to save HOT theories from common objections.

There seems to be a clear distinction between Rosenthal's HOT theory of consciousness and P-consciousness and therefore one cannot take Rosenthal's theory to serve as a basis for explaining the fundamentals of conscious experience. So far no problem.

Block however suggests that if we enrich the notion of HOT, we might be able to save the theory. Here is how:

P-consciousness is phylogenetically much older than any higher order cognition. Therefore in order for HOT to bridge the gap between higher level cognition and phenomenality, HOT has to accomodate P-consciousness. Here is a quote from Block's handout: "S is an R-conscious state iff S is phenomenally represented in a thought about S."

Just to remind you that R consciosness is in Block's terms just equivalent to HOT. Is this gonna move us anywhere? I have some doubts about it. Let us forget for now my general worry as to what exactly is meant here by "being phenomenally represented"

The issue which puzzles me in this respect is a following one. Let us take Rosenthal's example of Tip-of-Tongue experience. I am afraid that introduction of phenomenality at this stage is going to produce more harm

than benefit. I grant that all conscious beliefs have some phenomenality. But is the phenomenality in the case of different TOTs also different?

Do I experience the same thing as I am having TOT answering question WHAT IS THE CAPITAL OF FRANCE as opposed to WHAT IS THE DIAMETER OF THE EARTH? It seems like I have to, because my access goes to different ideas (although it does not reach their content). But at least to me, the feeling is just about the same - I think I know the answer, but I just can't produce it now.

So it seems like we might face a case of (potentially infinite) cases of access consciousness accompanied with a single experience. But that contradicts all our effort to keep them together...

Surely one can reply by saying that in fact all my individual TOTs are different (in the first case I feel that I know what is the capital of France, but I just can't say it, other time I feel like I know what the diameter is, but just can't say it...)

But what if my phenomenality does not go that far and I only feel that I know the answer, but I am just stuck.

At least it is an interesting open possibility...

Juraj

From jmartine@U.Arizona.EDU Sun Oct 31 17:00:11 1999
Return-Path: <jmartine@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA06960
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 17:00:11 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHSIHGNWTC9GZLPV@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
31 Oct 1999 17:00:58 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHSIHG4DB49JHPUD@Telcom.Arizona.EDU>
for chalmers@Ariziona.EDU; Sun, 31 Oct 1999 17:00:58 -0700 (MST)
Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA06954 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 17:00:01 -0700
Received: from localhost (jmartine@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id RAA24593; Sun,
31 Oct 1999 17:00:49 -0700 (MST)
Date: Sun, 31 Oct 1999 17:00:49 -0700 (MST)
From: Joel A Martinez <jmartine@U.Arizona.EDU>
Subject: Re: Block and HOT
In-reply-to: <Pine.A41.4.10.9910311455020.42206-100000@f1n2.u.arizona.edu>
To: Juraj Hvorecky <hvorecky@U.Arizona.EDU>
Cc: Science of Consciousness <scicon@paradox.soc-sci.arizona.edu>
Message-id: <Pine.HPX.4.10.9910311544230.24153-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Juraj raises an interesting point. It seems like Rosenthal does have to hold that the phenomenality in the case of different TOTs is different. In fact, he does think this. He states in "Consciousness and Metacognition"

There is a vivid conscious difference between having one word on the tip of one's tongue and having another. We are conscious of the informational state somehow without being conscious of the information it bears.

I too have trouble understanding exactly what Rosenthal means here. What is the phenomenality of a TOT? Perhaps his discussion of the difference between being conscious of a particular informational state and being conscious of the same state *in the relevant way* may be of help here.

Rosenthal uses the example of having Mark Twain's real name on the tip of one's tongue. The feeling of <knowing *what* Mark Twain's real name is> is different from the feeling of <knowing *that* Mark Twain's real name is Samuel Clemens>.

Is it possible to extend this sort of example to two different TOTs? I don't see why not. Wouldn't the feeling of knowing <what the capital of France is> be phenomenally different from the feeling of knowing <what the diameter of the earth is>? I think so. If that is right then Rosenthal is not committed to a number of different A-conscious states having the same phenomenology; something which Juraj was worried about.

On second thought, I am not sure why I think there is a phenomenological difference here. I suppose all I am really doing in this posting is presenting, in Rosenthal's terms, the opposing intuition to Juraj's. But, there might be more of an argument here. If there is a difference in phenomenology between TOT's I submit that it arises from a difference in description. That is, in TOT-A the relevant informational state bears the description <the capital of France> in TOT-B the relevant informational state bears the description <the diameter of the earth>. Is this enough to give us different phenomenology?

If you hold that all conscious beliefs have some phenomenology, then you need to explain what accounts for the phenomenological difference between two different conscious beliefs. It seems that description could play a role here. That is, two different conscious belief states have different phenomenology just in case they contain different descriptions. If one buys this claim for any conscious belief, it seems to me that one would have to buy it in the case of TOTs. For, TOT-A and TOT-B contain two different descriptions.

(Note: there may be a counter example to my account. Consider the following two TOTs:

- 1) <knowing the name of the planet to which "the evening star" refers.>
- 2) <knowing the name of the planet to which "the morning star" refers.>

The above two TOTs contain different descriptions. However, it may seem odd to say that the phenomenality is different, since in both cases one has "Venus" on the tip of one's tongue. I am not sure what to say about this yet, since I just thought of it.)

-Joel

Joel A. Martinez
Social Science Bldg. Rm. 213
Dept. of Philosophy
University of Arizona
P.O. Box 210027
Tucson, AZ 85721-0027
USA

Office #138 Phone 520-621-7098

All things excellent are as difficult as they are rare.

-Spinoza, The Ethics

From logant@U.Arizona.EDU Sun Oct 31 17:49:51 1999
Return-Path: <logant@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA07147
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 17:49:50 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHSK81M4GG9GZM5D@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
31 Oct 1999 17:50:39 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHSK814ALS9JFQRF@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 31 Oct 1999 17:50:38 -0700 (MST)
Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id RAA07142 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 17:49:42 -0700
Received: from localhost (logant@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id RAA05784 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 17:50:31 -0700 (MST)
Date: Sun, 31 Oct 1999 17:50:31 -0700 (MST)
From: Logan T Trujillo <logant@U.Arizona.EDU>
Subject: HOTS, unconscious processes, and qualia
In-reply-to: <Pine.HPX.4.10.9910311544230.24153-100000@orion.U.Arizona.EDU>
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910311705480.25353-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Hello all.

I was thinking about the discussion in class concerning the distinction between a conscious mental state and being conscious of a mental state. Dave gave a good example of this distinction (I am describing this from memory, so if I have this example wrong please let me know): show a subject a visual stimulus consisting of a large number of dots for a brief period of time, and then follow this with a second presentation of a group of dots whose number is one less than that of the original presentation. In most cases people would be unaware that there is a difference between the two sets of stimuli, and thus would report that the two sets are the same. It was argued in class that during the first stimulus presentation the extra dot was present in the person's phenomenology (i.e. the visual state corresponding to perception of the dot was conscious) but that the person was not aware of it (i.e. the person was not conscious of the dot).

Now it seems to me (using Block's terminology) that this implies a dissociation between phenomenal consciousness and access consciousness (or at least higher order thoughts that are conscious). One may say that while the subjects have phenomenal consciousness of the dot, they do not have access consciousness of the dot. The dot produces a shift or change in the subjects' phenomenology, but the subjects are not aware of this change because they don't have access to, or cannot discriminate, the change. But if one says this, then couldn't one also say that, in a sense, the experience of the dot is unconscious (or to be more precise un-access conscious)? If one can say this, then it leads naturally to the following questions: is it possible that all the mental processes

that are currently considered to be unconscious or nonconscious could be in fact conscious, but we have no access to them? Could these processes have experiential aspects that remain un-access conscious?

If one wished to deny this, how could one build a case from the evidence? Consider the case of a cognitive task where one wishes to infer that some unconscious process played a role in the instantiation of that task because the subjects undergoing the task had no awareness of the processes themselves. However one could equally say that the process is conscious (in a phenomenological sense) but that the subjects have no consciousness of the process (i.e. the process is un-access conscious). How could one empirically distinguish between these two interpretations?

In light of the strong arguments for the existence of the hard problem, it seems to me that the latter interpretation may be more metaphysically parsimonious. Instead of trying to figure out how qualia arises from certain brain processes while not arising from others, one posits that qualia arises from all brain processes. However, not all of these brain processes are in the category of which one is conscious of those processes. Of course this requires an explanation of why some processes make it into conscious access and others do not. However this may be a more tractable problem than the one suggested by denying that qualia is ubiquitous to all brain processes.

Take a look around you right now at your present visual scene. All the processes that enabled you to perceive that scene may contribute qualia to your overall visual phenomenology. The fact that we are not conscious of most of those processes does not argue against the possibility that such processes are conscious, with a rich phenomenology interwoven with the phenomenology of those processes to which we may be conscious of.

Logan T.

From chalmers@paradox.soc-sci.arizona.edu Sun Oct 31 19:40:37 1999
Return-Path: <chalmers@arizona.edu>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id TAA07636
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 19:40:37 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.2-31 #39830)
id <01JHS03CMVDS9GZ400@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
31 Oct 1999 19:41:24 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDf V5.2-31 #39830) with ESMTP id <01JHS03C8TZ49JHX62@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 31 Oct 1999 19:41:23 -0700 (MST)
Received: (from chalmers@localhost) by paradox.soc-sci.arizona.edu
(8.9.3/8.9.3) id TAA07626 for scicon; Sun, 31 Oct 1999 19:40:14 -0700
Date: Sun, 31 Oct 1999 19:40:14 -0700
From: David Chalmers <chalmers@arizona.edu>
Subject: HOTs, etc
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <199911010240.TAA07626@paradox.soc-sci.arizona.edu>
Status: R

Hi, a couple of notes:

(1) In response to Joel, one might try to make a case that the actual information that is on the tip of one's tongue can make a difference in phenomenology. E.g., if one is trying to think of the capital of Wyoming, it might feel different to have Cheyenne on the tip of one's tongue compared to having Laramie there. After all, there is often the feeling that some of the information is vaguely present -- maybe

the first letter of the world, or some of the feel. On the other hand, this feeling might be fairly coarse-grained, so it might be that e.g. Casper and Canter would feel the same?

(2) Logan suggests that in the dot case, the extra dot seems to be phenomenally conscious but not access conscious. I think one could make a case that we are access conscious of the extra dot -- it's visually there and available to play a role in our reasoning (e.g. counting the dots) and action (e.g. pointing to it). What it seems that we're not access conscious of is the fact that it *is* an extra dot -- i.e. the fact that the two displays contain different numbers of dots. But that's a slightly different thing.

The idea that lots of our supposedly "unconscious" mental states might be phenomenally conscious but access unconscious is interesting. But I don't think it gets support from this analogy, since we can point to and talk about the dot in the case above, but we can't point to and talk about typical "unconscious" information.

The motivation of suggesting that qualia are associated with all brain processes is interesting; but note that that doesn't force us into holding that they are all phenomenally conscious for me (the person with the brain). Another alternative might be that those qualia are not part of my phenomenal consciousness, but are part of another subject's consciousness (e.g., a proto-subject corresponding to an individual neuron, or neural system). Given that we're going to have qualia or proto-qualia all over, maybe it's not so bad to have subjects or proto-subjects all over, too. And this way we avoid the consequence that we as subjects experience a whole lot of qualia that we can never know about.

--Dave.

From acboch@U.Arizona.EDU Mon Nov 1 05:15:16 1999
Return-Path: <acboch@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id FAA09116
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 1 Nov 1999 05:15:16 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHT85U2B689GZS90@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Mon,
1 Nov 1999 05:16:04 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHT85TOYE09JHYW1@Telcom.Arizona.EDU>
for chalmers@Ariziona.EDU; Mon, 01 Nov 1999 05:16:03 -0700 (MST)
Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id FAA09111 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 01 Nov 1999 05:15:00 -0700
Received: from localhost (acboch@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id FAA26173; Mon,
01 Nov 1999 05:15:49 -0700 (MST)
Date: Mon, 01 Nov 1999 05:15:49 -0700 (MST)
From: Adam C Boch <acboch@U.Arizona.EDU>
Subject: Re: HOTs and Consciousness
In-reply-to: <19991031211141.11136.qmail@hotmail.com>
To: Matt Herbert <landsurveyor@hotmail.com>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9911010330130.21844-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

Matt writes:

>
> Yes, Rosenthal (almost) says that children have the HOTs about experience
> required for consciousness. A few lines after the passage cited by Peter,
> Rosenthal says that the essential difference is between thought and report.
> Even adults have thoughts that they cannot formulate as reports, so
> presumably, children do too. Some such thoughts may be HOTs-- "So it may
> well be that these children [three and younger] have the relevant HOTs even
> though they do not express them verbally."
>
> This line of thought has the interesting consequence of committing one to
> nativism about mentalese if one is to maintain that infants are conscious.
> If there's conscious experience, after all, it is so because it bears the
> imprint of a thought, and a thought implicates at least one concept.
> Regarding developing children who have not yet gained a large conceptual
> repertoire (expressed in a natural language), nativism might be invoked to
> explain the apparently rich conscious experience that children enjoy. It
> would not be *required* to explain consciousness, but it does seem like the
> natural direction to go if one is committed to nativism for infants.

I took it that Rosenthal wanted to go in a different direction. That is,
it seems that rather than say that children under three have a mentalese
(which, as an aside, does not commit one to nativism with respect to
mentalese since it seems *prima facie* plausible that a mental natural language
could form after birth and before verbalization) Rosenthal wants to say
that the relevant concepts necessary for a consciousness-bestowing HOT (a
concept of self and of mental states (mental state types and mental states
in general)) are merely content based. In other words, instead of a HOT
with the content "I am in M." where M is some mental state and where the
sentence "I am in M" is considered a mentalese sentence, the concept "I"
and "M" are minimalistic concepts filled out by a pre-linguistic content.

Exactly how this is to be cashed out is presumably left up to us since
Rosenthal doesn't give us much of a story. In "A Theory of Consciousness"
he says that

[a] minimal concept of self will suffice; no more than a
concept that allows distinguishing between oneself and other things...
Nor is there reason to suppose that rich conceptual resources are
necessary for a thought to refer to one's own mental states. We
refer in thought to physical objects by way of their position
in our visual field. It is natural to suppose that a thought can
similarly refer to sensory states by way of their position in the
relevant sensory field.

So it seems that here he wants to put forth some sort of demonstrative
reference story where a HOT merely points to a mental state (here a sensory
mental state). Mental state concepts (at least sensory ones) are
differentiated merely with respect to their content (e.g., a sensory
mental states S with the (phenomenal) content "reddish sphere" is
different from a sensory mental states M with (phenomenal) content
"greenish cube"). Since Rosenthal thinks one needs only a minimal concept
of self (presumably in the background of all thought-pointing events--as
Brad pointed out in class) children under three have all that they need to
form the relevant HOTs necessary for making mental states conscious
without the need to posit an infant pre-verbalized mentalese.

HOTs, then, for children under three, are, according to Rosenthal,
merely pointers to mental states. But, as Dave stated in class, it seems
Rosenthal is going to have to employ HOTs in this role even in adults.

Rosenthal writes (in ATOC):

In any case, conscious differentiation of sensory detail quickly outstrips one's conceptual resources; so some such means of referring to sensory states is necessary.

In other words, even as adults our sensory concepts are incommensurate with our sensory consciousness. We are conscious of sensory detail of which we have no corresponding sensory concept. So the HOTs in virtue of which we are conscious of our sensory mental states must "fill-in" by means of this demonstrative referring to which Rosenthal alludes.

There is, however, more story-telling to be told. For instance, if it is the richness of our mental state concepts that determines the richness of our conscious experience when HOTs are construed as mentalese sentences (which it seems Rosenthal is committed to claiming) then what is it that determines the richness of the "gaps" that are filled-in by these referring HOTs? Or, as is more glaringly problematic, what determines the richness of the conscious lives of children under three (since all of their HOTs are of this referring kind)?

From logant@U.Arizona.EDU Mon Nov 1 22:40:10 1999
Return-Path: <logant@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA11111
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 1 Nov 1999 22:40:09 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHU8NDNBBK9GZ0ZH@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Mon,
1 Nov 1999 22:41:00 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JHU8NDAJ3K9JIF4Z@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 01 Nov 1999 22:40:59 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA11093 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 01 Nov 1999 22:40:02 -0700
Received: from localhost (logant@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id WAA05161; Mon,
01 Nov 1999 22:40:52 -0700 (MST)
Date: Mon, 01 Nov 1999 22:40:51 -0700 (MST)
From: Logan T Trujillo <logant@U.Arizona.EDU>
Subject: Re: HOTs, etc
In-reply-to: <199911010240.TAA07626@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9911012205050.3398-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

On Sun, 31 Oct 1999, David Chalmers wrote:

> (2) Logan suggests that in the dot case, the extra dot seems to be
> phenomenally conscious but not access conscious. I think one could
> make a case that we are access conscious of the extra dot -- it's
> visually there and available to play a role in our reasoning (e.g.
> counting the dots) and action (e.g. pointing to it). What it seems
> that we're not access conscious of is the fact that it *is* an extra
> dot -- i.e. the fact that the two displays contain different numbers
> of dots. But that's a slightly different thing.

Perhaps the idea of access consciousness was the wrong thing to use to contrast against the phenomenality in this case, but it does seem that there is a distinction to be made. We can only point to and talk about the dot if there is an opportunity for our attention to make the dot's existence plain to us. Yes the dot is available for reasoning, etc., but such processes do not occur until conscious access to the dot is instantiated. Until then, the dot remains phenomenally present, but not accessed. Perhaps there could be a distinction between potential for access and actual access? Certainly the former obtains in this example, but does the latter? (I am admittedly out on a limb here).

>
> The idea that lots of our supposedly "unconscious" mental states might
> be phenomenally conscious but access unconscious is interesting. But
> I don't think it gets support from this analogy, since we can point to
> and talk about the dot in the case above, but we can't point to and
> talk about typical "unconscious" information.

Again, per the discussion above, I am not entirely convinced that the dot is access conscious before it is actually accessed. Thus the fact that we have the potential to point to the dot is not directly relevant to the current state of consciousness of actual access during the period before this occurs.

Furthermore wouldn't evidence from biofeedback and studies of meditation that purport to demonstrate effects of conscious intention upon autonomic physiology (processes usually considered to be unconscious) seem to indicate that in the case of some unconscious processes we can "point and talk" to them (metaphorically)? Perhaps states of consciousness exist where access is available to other typical unconscious processes? Such states certainly wouldn't be considered normal consciousness, but then normal conscious states are considered as such because the majority of people tested exhibit the inability to access information considered unconscious.

>
> The motivation of suggesting that qualia are associated with all brain
> processes is interesting; but note that that doesn't force us into
> holding that they are all phenomenally conscious for me (the person
> with the brain). Another alternative might be that those qualia are
> not part of my phenomenal consciousness, but are part of another
> subject's consciousness (e.g., a proto-subject corresponding to an
> individual neuron, or neural system). Given that we're going to have
> qualia or proto-qualia all over, maybe it's not so bad to have
> subjects or proto-subjects all over, too. And this way we avoid the
> consequence that we as subjects experience a whole lot of qualia that
> we can never know about.

This is a very interesting idea, perhaps closer to the idea I was trying to get across in my post. I think that such a situation is intimately associated with the problem of phenomenal/perceptual/cognitive binding. Perhaps binding regulates what qualia is associated with which subject. However whether or not this could be empirically tested remains to be seen.

Thanks Dave. As usual your comments and observations are astute and clarifying.

Logan T.

From anhabib@U.Arizona.EDU Sun Oct 31 21:59:04 1999

Return-Path: <anhabib@U.Arizona.EDU>

Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA08191

for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 21:59:03 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)

id <01JHSSX0V6U89GZ3N5@Telcom.Arizona.EDU> for

chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,

31 Oct 1999 21:59:51 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU

(PMDF V5.2-31 #39830) with ESMTP id <01JHSSX0MAN49JHZT5@Telcom.Arizona.EDU>

for chalmers@Arizona.EDU; Sun, 31 Oct 1999 21:59:51 -0700 (MST)

Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])

by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA08186 for

<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 21:58:54 -0700

Received: from localhost (anhabib@localhost)

by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id VAA05881 for

<scicon@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 21:59:42 -0700 (MST)

Date: Sun, 31 Oct 1999 21:59:42 -0700 (MST)

From: Allen N Habib <anhabib@U.Arizona.EDU>

Subject: Joel, Rosenthal and the HOT theory

To: scicon@paradox.soc-sci.arizona.edu

Message-id: <Pine.HPX.4.10.9910312138210.24318-100000@orion.U.Arizona.EDU>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: RO

All

Joel's post got me thinking about the need for HOTs to make sensory data conscious. Perhaps we can modify Rosenthal's proposal to make it more acceptable in the following manner: we remove the requirement that HOTs are necessary to make incoming sens data conscious, but are required for all other conscious thoughts.

Allow me to present an analogy to make my point. Imagine that memories are like data files. In order to become conscious, a memory must be 'copied' into the cache (working memory). The HOT is the equivalent of the copied file in the cache. Now files come in two parts, an index and a content part. The index segment contains information about the file, things like (in a computer) its size and origin and type. Perhaps our memories also contain an index, and when we have a TOT experience, we only have the index (or maybe even part of the index) copied into the cache, and that is why we have only partial consciousness. Only when the content part of the file is copied into memory will we be fully conscious of it. Of course, the problem is that, even on this generous analogy, sense data will not seem to require a HOT to be conscious, because the computer equivalent of sense data, data from peripherals that lie outside the bus (the information exchange, or CNS of a computer) is not copied into the cache via a subroutine, but is rather streamed into the cache through the bus. Perhaps this is what happens with sense data, then. Rather than eliciting an HOT, like other conscious experiences, sensory experiences would have a free ride to consciousness, as it were.

Although, this limitation might not be as devastating as it seems, since Rosenthal might argue that attending to sense data inevitably results in HOTs. Perhaps if we concentrate on a stream of incoming sense data, we, as a matter of empirical fact, begin to process the stream in a manner that requires the use of memory, and thus the production of HOTs. It is not implausible that the simple act of turning our attention to an incoming sensory stream results in our processing the units of the stream for memory matches, or trying to anticipate the upcoming units in the stream, or comparing past units looking for patterns, or any number of processes that would require HOTs on this formulation. So the HOT theorist can say that HOTs are necessary for a vast number of conscious experiences, including most of those inside attention. Of course, this argument is offered in the spirit of science fiction, and I don't have a good reason to think that human cognition is sufficiently similar to

computer architecture as to allow the inferences I have drawn here. . .
still, a good story, no?

From switanek@U.Arizona.EDU Sun Oct 31 22:31:20 1999
Return-Path: <switanek@U.Arizona.EDU>
Received: from Arizona.EDU (Penny.Telcom.arizona.edu [128.196.128.217])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id WAA08273
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 31 Oct 1999 22:31:20 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHSU20V8LC9GZMWP@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
31 Oct 1999 22:32:08 MST
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU)
by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
with ESMTP id <01JHSU20EUI89JI04R@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Sun, 31 Oct 1999 22:32:07 -0700 (MST)
Received: from localhost (switanek@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id WAA14243; Sun,
31 Oct 1999 22:32:06 -0700 (MST)
Date: Sun, 31 Oct 1999 22:32:06 -0700 (MST)
From: Nicholas J Switanek <switanek@U.Arizona.EDU>
Subject: Re: HOTs, etc
In-reply-to: <199911010240.TAA07626@paradox.soc-sci.arizona.edu>
To: David Chalmers <chalmers@Arizona.EDU>
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9910312107130.16516-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Juraj's post regarding the homogeneity of TOT phenomenology was interesting. There does seem to be significant commonality between all experiences of this. There is something distinctive about the brain stalling out because of (what I think might be described as) informational inadequacy. And Juraj's suggestion that the phenomenality of every "brain stall" is roughly the same is both intuitive and helpful. Yet it is hard to deny that tip-of-the-tongue experiences are strongly intentional. If we separate the TOT experience from the information a quest for which resulted in the TOT, and call all TOT phenomenologically equivalent, then it seems *any* answer to the frustrated search would dissolve the block, give that distinctive catharsis (and post-catharsis self-reprimand for having forgotten.) This is manifestly not the case, so it seems that paralyzing TOTs must need be chained to their antidotes, the informational states that are the TOTs' resolutions.

But to say the question, the quest, and the answer are linked is not yet to describe the phenomenological nature of the link. A TOT is like being lost. You've departed with a set of directions, but the directions are inadequate. There is a phenomenality that is specific to being lost with *this* set of directions (which corresponds to Joel's "description" and to Dave's "course-grained" "feel"), but the overwhelming feeling is simply that of being lost (which corresponds to Juraj's intuition that TOT phenomenology is all roughly the same). There is a sort of access consciousness, but what is available for verbal report is only that the access failed, the destination was not reached--along with some vague, "course-grained" information about where the destination "should be." The phenomenology of reaching the destination, of remembering the answer, is very little like that of being lost; the catharsis that comes with the answer arises because of this large disparity.

Or else a TOT is like having an adequate set of directions, but an outdated one. A metaphor that seems more apt is getting lost when driving home after a long time away and after the surroundings have changed. I'm originally from Phoenix, which has been spreading, mutating, and losangelizing for the past couple of years: eight months is time enough for any number of gaudy strip malls to swell like lumps on the desert. So when I meant to drive home after a stay abroad, and looked for the old clues, I was disoriented by the faux adobe franchises. The city planning of the mind is also in flux (please correct me if I'm wrong on this) which can disorient quests for information, apart from any consideration of the quality of the set of directions, and whether the ink has faded or the page has torn.

Last, the getting lost metaphor can be extended to be suggestive in the case of overconfident feeling-of-knowledge. The executive "manages" all the business of the mind with the outside. He takes the request for a memory and sets his people working. If the only answer his consulting firm is able to come up with is that they can't get the requested information, but they can describe some characteristics of it, the executive, in order to save face and continue to legitimize his high fees, claims he's eighty percent sure of his guess. Or else the executive is a nimwit who doesn't understand what his techies and sundry minions do in their cubicles, but is confident and dependent upon them, regardless of their troubles--indeed, he can't understand their troubles. Whatever the case, the executive is clearly male, and hates getting lost.

Nick

From rachaelp@U.Arizona.EDU Tue Nov 2 07:45:12 1999
Return-Path: <rachaelp@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.arizona.edu [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id HAA11844
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 2 Nov 1999 07:45:12 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.2-31 #39830)
id <01JHUR05YVY89H01ND@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Tue,
2 Nov 1999 07:46:03 MST
Received: from odo.U.Arizona.EDU by Telcom.Arizona.EDU (PMDf V5.2-31 #39830)
with ESMTP id <01JHUR05NTI89JIGGI@Telcom.Arizona.EDU> for
chalmers@Ariziona.EDU; Tue, 02 Nov 1999 07:46:03 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-1.U.Arizona.EDU [128.196.137.195])
by odo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id HAA00106 for
<chalmers@odo.U.Arizona.EDU>; Tue, 02 Nov 1999 07:45:55 -0700 (MST)
Received: from localhost (rachaelp@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id HAA01819 for
<chalmers@U.Arizona.EDU>; Tue, 02 Nov 1999 07:45:55 -0700 (MST)
Date: Tue, 02 Nov 1999 07:45:54 -0700 (MST)
From: Rachael J Parkinson <rachaelp@U.Arizona.EDU>
Subject: consciousness and meta-processing
In-reply-to: <Pine.HPX.4.10.9910251308231.19303-100000@pavo.U.Arizona.EDU>
To: David Chalmers <chalmers@U.Arizona.EDU>
Message-id: <Pine.HPX.4.10.9910281124390.25003-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Dear Dave,
I wrote this message last Thursday. I thought that I had sent it out but instead I had hit 'postpone'. This whole time I have been thinking that I had posted to the listserve with plenty of time. (happy

that I hadn't procrastinated until Sunday.) When I was talking with Adam last night I figured out that he hadn't gotten my post and that further, it hadn't been sent. So here it is. Sorry that it is late but know that I did finish writing it on time. If you decide to send it on to the listserve please take off this excuse (its an embarrassing mistake to make.)

Thanks, Rachael

Our discussion on Tuesday raised another concern that I had with the Johnson and Reeder article. In the section, Consciousness and Attention, on page 283, they make some pretty interesting claims about consciousness and attention. In this section, they say that the familiar example of driving the car to work is one where the person is paying attention but is not necessarily conscious. This is because consciousness can be distinguished from attention "in that consciousness includes the idea that processes are engaged and monitored by another subsystem, whereas attention does not necessarily involve transactions across subsystems."

So it appears that Johnson and Reeder want to claim that what is going on when we are 'unconsciously' driving to work is that transactions are taking place across one subsystem but processes are not being engaged and monitored across subsystems. But recall, theirs is a picture in which the fulfilling of agendas requires control and monitoring *across* subsystems. (Strangely, the MEM framework doesn't leave room for these functions within a particular subsystem). So what they would have to say about the experience of driving 'unconsciously', while 'paying attention' is one in which no agendas are being carried out, that is, where no control or monitoring is taking place. Anyone who has driven to the store, even 'unconsciously' should recognize that some controlling and monitoring must take place as accomplishing this task requires heading in a basic direction, changing lanes, and in general, not being killed. Even on 'automatic' there must be some control/monitoring processes taking place for us to accomplish our goal.

There are a number of ways Johnson and Reeder could adjust their position to make it more tenable. First, as I pointed to before, they could allow for control and monitoring to take place across a particular subsystem. Secondly, as Liz pointed out, their description of the driving example seems to get things backwards. Most of our intuitions seem to suggest that we are conscious but not paying attention in those scenarios, not the other way around. I am not sure that this claim about consciousness and attention coming apart is that important for Johnson and Reeder's overall project anyway. The claim that consciousness arises from meta-processing (though problematic in itself) is better supported without tacking it to the stronger claim about what attention involves.

Just some thoughts...

From bayne@U.Arizona.EDU Wed Nov 3 13:54:05 1999
Return-Path: <bayne@U.Arizona.EDU>
Received: from Arizona.EDU (Hopey.Telcom.Arizona.EDU [128.196.128.234])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id NAA15222
for <chalmers@paradox.soc-sci.arizona.edu>; Wed, 3 Nov 1999 13:54:05 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JHWIUUVW1PS9H00KF@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Wed,
3 Nov 1999 13:54:58 MST
Received: from f1n7.u.arizona.edu by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
with ESMTP id <01JHWIUUVITKG9JIXFQ@Telcom.Arizona.EDU> for
chalmers@Arizona.EDU; Wed, 03 Nov 1999 13:54:57 -0700 (MST)
Received: from localhost (bayne@localhost) by f1n7.u.arizona.edu (8.8.8/8.8.8)
with ESMTP id NAA42498; Wed, 03 Nov 1999 13:54:57 -0700
Date: Wed, 03 Nov 1999 13:54:57 -0700 (MST)
From: Timothy J Bayne <bayne@U.Arizona.EDU>
Subject: Nagel
In-reply-to: <Pine.HPX.4.10.9910312107130.16516-1000000@pavo.U.Arizona.EDU>
Cc: David Chalmers <chalmers@Arizona.EDU>, scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9911031339520.31990-1000000@f1n7.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

I had a discussion with someone this morning about Nagel and some interesting stuff seemed to come out of it. I haven't reread Nagel, so it may not be true to his intent.

(1) Think of an objective phenomenology as something like an (allocentric) space-time map of the world. Every physical thing in the world is at some space-time distance from every other thing. Distinct beings can share this conception of the world. Similarly, a phenomenological dimensional space is a framework onto which all (possible?) phenomenal states can be mapped. Our own (actual and possible) experiences fall somewhere in this multi-dimensional space. The bat's experiences will occur somewhere within this space too.

(2) First problem: what is the structure of this phenomenal space? At least two parts to that problem: (a) what is the structure of our own phenomenal space? (b) how does the structure of our phenomenal experience relate to that of any possible phenomenal experience?

(3) Second problem: where on this phenomenal map does the bat fall? Tough one, but interestingly distinct from (2).

(4) First thesis: having experience x is not sufficient for knowing what it's like to experience x. Knowing what it's like to experience x involves understanding the dimensional structure of x, but one can have x without appreciating this dimensional structure. A being that only experienced a single shade of blue might have no appreciation of the structure of color.

(5) Second thesis: Having an understanding of some structural aspect of experience x might enable you to partially know what it's like to experience x. The sound of trumpets and red might share a high-level structural feature, which makes this analogy partially illuminating.

(6) Third thesis: The richer your understanding of the structure of an experience, the closer you can get to understanding it without having it. Thus, Hume on the missing shade of blue: if you have experienced the other shades, then you probably have an understanding of most of the dimensions that will structure the missing shade.

(7) Fourth thesis: An objective phenomenology enables you to go from understanding what it's like to experience x to understanding what it's like to experience anything else (perhaps given that there's some structural similarity), but it doesn't bootstrap you into understanding phenomenal states from an absence of phenomenal states. That is, if we had this objective phenomenal map, we could understand what it's like to be a bat (or anything else), but zombies couldn't use this map, 'cos they can't get positioned on it at all. Objective phenomenology allows you to get from one phenomenal state to others, it doesn't allow you to go from merely physical states to phenomenal states. Thus, Mary can use it to know what it's like to see red (without seeing red), but Zombie Mary can't.

tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From lnielsen@azstarnet.com Sun Nov 7 14:51:24 1999
Return-Path: <lnielsen@azstarnet.com>
Received: from Arizona.EDU (Hopey.Telcom.Arizona.EDU [128.196.128.234])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id OAA20575
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 7 Nov 1999 14:51:24 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JI261G2Z1S9H0QSV@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
7 Nov 1999 14:52:23 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JI261FPP289JJCYE@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 07 Nov 1999 14:52:22 -0700 (MST)
Received: from cepheus.azstarnet.com (cepheus.azstarnet.com [169.197.56.195])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id OAA20571 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 07 Nov 1999 14:51:15 -0700
Received: from zippo (dialup002ip475.tus.azstarnet.com [169.197.15.219])
by cepheus.azstarnet.com (8.9.3+blt.Beta0/8.9.3) with SMTP id OAA28891 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 07 Nov 1999 14:52:11 -0700 (MST)
Date: Sun, 07 Nov 1999 14:58:14 -0700
From: Lis Nielsen <lnielsen@azstarnet.com>
Subject: Animal Consciousness
To: scicon@paradox.soc-sci.arizona.edu
Reply-to: lnielsen@u.arizona.edu
Message-id: <NCBBJCIPCKAPOGDICDPAEJOCDA.lnielsen@azstarnet.com>
MIME-version: 1.0
X-MIMEOLE: Produced By Microsoft MimeOLE V5.00.2314.1300
X-Mailer: Microsoft Outlook IMO, Build 9.0.2416 (9.0.2910.0)
Content-type: text/plain; charset=iso-8859-1
Content-transfer-encoding: 8BIT
Importance: Normal
X-Priority: 3 (Normal)
X-MSMail-priority: Normal
X-Sent-via: StarNet <http://www.azstarnet.com/>
Status: RO

Part 1. Gallup is confused, and it's not about consciousness at all.

The term consciousness is used loosely in Gallup's article, and the real debate between Gallup and Povinelli concerns whether the animals they study have a theory of mind or 'empathy' (also rather loosely conceived). Specifically, do they have the sort of self-concepts that allow for understanding of their own mental states as mental states, and can they extend this knowledge to encompass an appreciation that other creatures like themselves have mental states like their own? The relationship between theory of mind and consciousness, according to Gallup, seems at first to go something like this: If animals have a theory of other minds, then they must first have a theory of their own minds, and having such a theory makes them likely possessors of consciousness (presumably because if they're thinking about all these mental states, they must be doing it consciously). Later, in some sweeping claims at the end of the article, Gallup equates consciousness and self-consciousness. He suggests that creatures without self-consciousness may have "clever brains but blank minds." Having a blank mind is more than not having a theory of other minds; it is compared to sleepwalking or blindsight, states in which sensory inputs can influence behavior, without any phenomenal consciousness of that input. Blank minds are really phenomenally empty. The link between self-consciousness and phenomenal consciousness turns out to be causal, according to Gallup, who (revealing his Cartesian allegiance) writes, "Being able to conceive of oneself in the first place is what makes consciousness and thinking possible."

Povinelli's statement of the focus of the debate is more clear, and he avoids the term consciousness altogether. He asks, "[W]hat excludes the possibility that evolution has simply produced "mind-blind" mechanisms that lead social primates to look where other animals look, without entertaining any ideas about their visual perspective?" Why ascribe a theory of mind to these creatures, he asks, when a more parsimonious explanation of their behaviors will suffice? He offers the idea of a kinesthetic self-concept to account for the behavior of animals in the mirror task, rejecting interpretations of this ability that infer an understanding of psychological states as psychological states on the animal's part. Povinelli doesn't seem to share Gallup's intuitions about the relationship between theory of mind and consciousness. What the animals lack, he says, is the ability to use their own experiences (phenomenally conscious, I presume) to make inferences about the experiences of others. Povinelli might be interpreted as sharing Nagel's assumption that animals have conscious experiences.

Part 2. But since Gallup brought it up, let's talk about it anyway.

What does this discussion reveal about debates about animal consciousness? In particular, what exactly are these investigators imagining when they talk about "blank minds" filled with "mind-blind" mechanisms? Are such minds phenomenally conscious, but merely lacking in a theory of other minds? Or lacking a theory of mind, are these animals not conscious at all, as Gallup suggests?

There are some fundamental assumptions lurking behind the idea that the burden of proof in discussions of animal consciousness lies with those who would argue for it, as opposed to those who would argue against. Morgan's Cannon, as we learned on Tuesday, cautions us to be parsimonious. We should refrain from offering complex explanations when more simple explanations suffice, therefore, we should not claim that consciousness exists in animals if simpler explanations can account for the observed behaviors. The assumption here is that there is an ascending hierarchy of mental capabilities or functions, with consciousness at the top. On Gallup's view we have consciousness, self-consciousness and consciousness of other minds stacked at the top. Beneath this top level are abilities like learning,

memory, and so on that can occur in blank minds. Phenomenal consciousness does not constitute a level of its own. For Povinelli, phenomenal consciousness may have a its own level, beneath the level where we find theory of mind and self-consciousness. Both investigators make the common assumption that many mental functions are mere mechanisms requiring no awareness for their function, mechanisms that inhabit lower levels of the hierarchy. Animal minds should be explained at the lowest level of the hierarchy that can account for the observed behaviors.

This highly mechanistic view of animal minds is a direct inheritance from certain recurrent themes in our western intellectual tradition. It has ancient roots (both in Greek philosophy and Christianity) in a form of speciesism that places humans at the top of the psychological and biological heap. In Descartes, this view takes the form of a mind-body dualism in which our mental nature is markedly different from and superior to our bodily nature. When cast in evolutionary terms it is the view that human consciousness is the most highly evolved mental function in the animal kingdom. Combined with the notion that animals don't have minds (or souls) at all, such views helped to justify the beginnings of animal experimentation. We essentially have two parallel hierarchies that are assumed to map onto one another. The biological hierarchy has humans at the top, followed by other primates, then 'lower animals' of all sorts. The psychological hierarchy has consciousness at the top (maybe self-consciousness above it), complex cognition (memory, learning, etc.) below, and behavior at the bottom. The assumptions here are that the higher levels are more complex. This assumption is incorporated into cognitive models of consciousness that place consciousness at the center or top of the information processing system, and in evolutionary explanations of consciousness that argue for its role in dealing with the demands of more complex information processing systems interacting with more complex environments.

In a recent article in the Journal of Consciousness Studies (1998, Volume 5, 3) entitled "Consciousness: A Natural History", Maxine Sheets-Johnstone questions the correctness of these hierarchical conceptions and of the assumption that "unconsciousness historically preceded consciousness" in animals. She suggests that proprioception may be the first evolved form of consciousness. The evolution of proprioception, she proposes, parallels the evolution of animate forms, such that from the very beginning of the ability of organisms to move, there was a need for a kind of flexible responsivity to external stimuli. It is arbitrary, she argues, to call this responsivity behavioral or cognitive when referring to 'lower animals' and conscious when referring to humans or 'higher animals.' The fact that this is frequently done has much to do, she claims, with our brain-centered notions of consciousness that disregard more embodied sensory abilities. She notes that the first human sense to develop is proprioception (it develops prenatally with the early development of motor pathways), and it is through this sense that we initially come to learn to move our bodies and to feel ourselves. This is a sense that we share with many 'simple' creatures. Sheets-Johnstone provides a description of the proprioceptive abilities of invertebrates that makes the assumption that unconscious mechanisms explain the behaviors of these 'lower animals' look disturbingly ad hoc. In response to Dennett, who claims that in simple organisms "there is really nothing much to self-knowledge beyond the rudimentary biological wisdom enshrined in such maxims as 'When Hungry, Don't Eat Yourself!' and 'When there's a Pain, It's Yours!'" she questions the parsimony of explaining animal behaviors in terms of such mechanisms. I'll close with the following quotation in which she makes this point:

"[W]e should ask what it means to say that a lobster will eat another's claws but that conveniently, as Dennett puts it, it finds eating one of its

own claws unthinkable. Does it mean that there is actually a rule 'Don't eat your own claws!' wired into the lobster's neurological circuitry? But it is patently unparsimonious to think that there is such a rule and just as patently absurd to think that every creature comes prepared with an owner's manual, as it were, a rulebook replete with what Dennett calls 'maxims'. Such a maxim, for example would be only one of an indefinitely great number of maxims that a lobster (or, in analogous terms, any other 'simpler organism') could be said to carry around in the neural machinery that counts as its 'Headquarters'; 'Don't try to go on land!' 'Don't try to eat a squid!' 'Shovel in new sand grains after molting!' 'The large claw is for crushing!' 'The small claw is for seizing and tearing!' And so on. ... What makes eating its own claws 'conveniently unthinkable' is clearly something other than a rule of conduct. 'Convenience' is not a matter of an opportune adaptation but of an astioundingly varied and intricately detailed biological faculty that allows a creature to know its own body and its own body in movement. ... These kinetic cognitional abilities constitute a corporeal consciousness. ... A moment's serious reflection ... discloses a major reason why ... sensitivity to movement is both basic and paramount: no matter what the particular world in which an animal lives, it is not an unchanging world. Hence, whatever the animal, its movement cannot be absolutely programmed such that, for example, at all times its particular speed and direction of movement, its every impulse and stirring, its every pause and stillness, run automatically on something akin to a lifetime tape" (Sheets-Johnstone, 1998, pp. 274-8).

Offering mechanistic explanations for animal behaviors may reveal more about one's commitment to certain assumptions about the mappings between certain presupposed biological and psychological hierarchies of complexity than it does about one's commitment to parsimony of explanation.

Lis Nielsen
Department of Psychology
University of Arizona
Tucson, AZ 85721-0068
USA

lnielsen@u.arizona.edu

From press@U.Arizona.EDU Sun Nov 7 16:52:56 1999
Return-Path: <press@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.Arizona.EDU [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA20655
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 7 Nov 1999 16:52:56 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JI2AA4BP3K9H1B93@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
7 Nov 1999 16:53:55 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JI2AA304FK9JJTXD@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 07 Nov 1999 16:53:54 -0700 (MST)
Received: from f1n2.u.arizona.edu
(IDENT:press@f1n2.U.Arizona.EDU [128.196.137.102])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id QAA20650 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 07 Nov 1999 16:52:46 -0700
Received: from localhost (press@localhost) by f1n2.u.arizona.edu (8.8.8/8.8.8)
with ESMTP id QAA33180 for <scicon@paradox.soc-sci.arizona.edu>; Sun,
07 Nov 1999 16:53:45 -0700
Date: Sun, 07 Nov 1999 16:53:45 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>

Subject: Mirror Test

To: scicon@paradox.soc-sci.arizona.edu

Message-id: <Pine.A41.4.10.9911071650370.34642-100000@f1n2.u.arizona.edu>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: RO

All,

Here are some thoughts regarding the hypothesis (as discussed during our last class) that a basic understanding of mirrors plus what we were calling a self-schema might allow chimps to pass the Mirror test just as well as a full-fledged self concept. Dave seemed intrigued by this idea, but rightly pointed out that it was unclear whether the difference between the two views is empirically detectable. During the break, Juraj and I tossed around a few ideas about how an empirical test might work, but didn't really get very far. After a bit more thought, though, I think something along the lines of what we were talking about might work.

To clarify the hypothesis, the idea is that mirrors give a sighted creature visual information about areas of space that are not in the direction the creature's eyes are pointing. In particular, they give us information about areas of space defined by Newton's laws of optics (angle of incidence = angle of reflection). Obviously, chimps don't understand optics, but they might be able to learn by trial and error how to coordinate mirror information with the rest of their visual information in such a way that they could locate objects seen in a mirror, grasp them, and so on. If so, locating the blue spot on their forehead is just a special case of using mirrors to locate things. Once the chimp realizes that the spot is located in a region of space that coincides with its skin, the self-schema can take over. That is, chimps may just be designed in such a way that they take a special interest in the region of space that coincides with their bodies without realizing that it is "their" body.

Falsification of this hypothesis doesn't seem to be a problem. Since the hypothesis sees success in the Mirror test as just a special case of using mirrors to locate things, if chimps can't generally use mirrors to locate things other than themselves, it would be clear that something else was involved in the Mirror test. We might not want to require that they be able to use mirrors at all angles, since arguably the information they get from a mirror at, say, 45 degrees to their line of sight is harder to interpret than the information they get from a mirror that's close to perpendicular to their line of sight. But if it turns out that a chimp can locate spots on his body using the mirror, but not the banana right next to him (hidden from direct line of sight) that would tend to disprove the claim that chimps have a general ability to understand the visual information in mirrors.

Confirming the hypothesis seems to be the hard part. The general trend in the ideas that Juraj and I discussed involved introducing various sorts of degraded visual conditions. The idea was that one of the two competing methods (self-concept vs. self-schema plus an understanding of mirrors) was likely to be harder to adapt to degraded conditions than the other. If so, then we might be able to figure out which method was being used from the way that performance dropped off as a function of the visual degradation. One way of degrading the visual conditions would be to introduce distorting mirrors (i.e. like the ones in fun-houses at carnivals).

Of course, the problem is that I have no idea which of these methods is more/less adaptable to these sorts of conditions. Juraj and I

tried to come up with a priori reasons to think that one task would be harder than the other, but didn't come up with anything convincing. The alternative I've come up with is to use human performance as a baseline for the relative difficulty of the adaptive tasks. Presumably, human beings have both a self-concept and a self-schema with mirror-understanding. If we can test these abilities separately in humans, we might get a quasi-objective measure of which type of process is more adaptable to fun-house mirrors. Then we test the chimps and compare their drop in performance to the way that performance dropped off in humans to see whether it was more like the drop off on the self-concept test or the self-schema test. Here's what I have in mind:

1. Human self-concept test. We can't use humans looking at their own reflections in distorted mirrors, because even the densest human understands mirrors well enough know that the distorted image is of himself, even if it is totally unrecognizable. Instead, we set up a hidden video camera right on top of a monitor which displays its output. Now the subject can't rely on his knowledge of the mechanism producing the image to identify it. Of course, as soon as he does identify it, he will realize that there is a camera somewhere, but by then we have our data. This task should be tried with varying amounts of distortion, preferably distortion that mimics the distorted mirrors used in the rest of the test.

2. Human mirror location test. Set the mirror up in such a way that the subject has to use the mirror to locate and grasp some object that he can't see directly. Again, we try this with increasingly distorted mirrors.

The data we get from these two trials should be a learning curve for each of the tasks. How long does it take the subjects to figure out how to use the new (increasingly) distorted mirrors to accomplish 1 and 2? Presumably, both tasks will get more difficult as the mirrors get more distorted. Hopefully, they will get more difficult in characteristically different ways. For example, one ability might decline geometrically, while the other declines linearly, or whatever. Then, it's the chimps' turn:

3. Give the Mirror test to the chimps using increasingly distorted mirrors. Again, their performance on the test should get worse as the test gets harder. If we are lucky, their learning curve for this task will strongly resemble one, but not the other, of the curves obtained from 1 or 2. We would then have at least some evidence that the chimp was using that method.

Obviously, there are some serious experimental difficulties (comparing the human and chimp data, getting curves that are clearly different enough to strongly support one hypothesis over the other, just to name a couple). The whole thing may just be unworkable, but I thought I'd throw it out in the open where other more experimentally minded folks might either find ways to improve it or put it out of its misery. Suggestions?

Experimentally,
Joel

From sawright@U.Arizona.EDU Sun Nov 7 21:50:51 1999
Return-Path: <sawright@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.Arizona.EDU [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA20819
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 7 Nov 1999 21:50:51 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JI2K0IE1SW9H01WC@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
7 Nov 1999 21:51:51 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JI2K0I2QM09JJEJD@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 07 Nov 1999 21:51:50 -0700 (MST)

Received: from f1n3.u.arizona.edu
(IDENT:sawright@f1n3.U.Arizona.EDU [128.196.137.103])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id VAA20815 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 07 Nov 1999 21:50:47 -0700

Received: from localhost (sawright@localhost)
by f1n3.u.arizona.edu (8.8.8/8.8.8) with ESMTP id VAA37540 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 07 Nov 1999 21:51:43 -0700

Date: Sun, 07 Nov 1999 21:51:43 -0700 (MST)
From: Sarah A Wright <sawright@U.Arizona.EDU>
Subject: Evolutionary explanation
To: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9911072148440.37800-100000@f1n3.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: RO

Since I had to miss out on the discussion of Tuesday, due to illness, I hope that these comments aren't too far off the discussion there. They are a bit tangential, but I think of general interest for assessing the evidential support of primate studies.

In the section "The Meaning of Self-Recognition" Povinelli ties his research together, and comes to the conclusion that great apes have a Kinesthetic self-concept, and not a robust psychological self-concept. After reaching this conclusion he tacks on an extra theory about the evolutionary source of a kinesthetic self-concept. My concern is that this add-on is not just another theory which is related and supported by the same evidence, but rather used as a support for Povinelli's primary conclusion. If Povinelli wants the evolutionary theory to play this supporting role, then I think that we must analyze the sort and strength of support that such explanations can give to a theory of current processing.

So we have a story according to which it is of benefit to the great apes to develop a kinesthetic self-concept. They got so big, so fast, maneuvering around in the treetops required a new self-regulation, and a kinesthetic self-concept can play the regulating role. Further it is all that is needed; a psychological self-concept does not help one swing from the trees any better. But what about gorillas? They are an anomaly, being both large, and failing the self-recognition tasks. Is this a counterexample? No, since gorillas are primarily land-dwelling. As land dwellers, they lost the need for fine-grained self-regulation, and lost the kinesthetic self concept. Thus Povinelli's theory can explain the source of the self- concept and its exceptions, and this seems like an additional virtue of the theory.

But is it? While kinesthetic self-regulation can be worked into a historical story, so can psychological self-regulation. Even if the simpler form of self-regulation will do for regulating tree swinging behavior, that does not count against possibility of a more complex form. Only if we accept a further claim (which is often not clearly articulated in these sorts of arguments) that natural selection will pick the minimal way of achieving an ability. To argue in this way is to ignore the real possibility of spandrels, or free-riders, in the evolutionary game. But

while certain theories play this possibility up or down, none deny that we can have an occasional spandrel; appeals to simplicity (particularly when the alternatives are close in complexity) can't be decisive.

While this objection doesn't undermine Povinelli's other arguments, I do think that the introduction of evolutionary explanations (here and elsewhere) can be done in such a way to suggest support without actually giving any. Perhaps support isn't Povinelli's aim in giving his evolutionary explanation. But I think in this case, and in others where support is suggested, we should be critical.

Any comments?

Sarah

From serobert@U.Arizona.EDU Sun Nov 7 23:40:14 1999

Return-Path: <serobert@U.Arizona.EDU>

Received: from Arizona.EDU (Hopey.Telcom.Arizona.EDU [128.196.128.234])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA21099
for <chalmers@paradox.soc-sci.arizona.edu>; Sun, 7 Nov 1999 23:40:14 -0700

Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JI20I40UHS9H1JGZ@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Sun,
7 Nov 1999 23:41:13 MST

Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JI20I3M9XC9JIZBD@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Sun, 07 Nov 1999 23:41:12 -0700 (MST)

Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id XAA21094 for
<scicon@paradox.soc-sci.arizona.edu>; Sun, 07 Nov 1999 23:40:07 -0700

Received: from localhost (serobert@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id XAA27243; Sun,
07 Nov 1999 23:41:07 -0700 (MST)

Date: Sun, 07 Nov 1999 23:41:07 -0700 (MST)

From: Simon E Roberts-Thomson <serobert@U.Arizona.EDU>

Subject: Povinelli

In-reply-to: <Pine.A41.4.10.9911072148440.37800-100000@f1n3.u.arizona.edu>

To: Sarah A Wright <sawright@U.Arizona.EDU>

Cc: scicon@paradox.soc-sci.arizona.edu

Message-id: <Pine.HPX.4.10.9911072250130.5757-100000@orion.U.Arizona.EDU>

MIME-version: 1.0

Content-type: TEXT/PLAIN; charset=US-ASCII

Status: R

I wanted to address some of Sarah's concerns. When Povinelli is talking about kinesthetic self-concept, it actually seems as though he has two arguments in mind, and it is not at all clear that they are compatible.

(2) Heavy Tree Dwellers vs Light Tree Dwellers.

According to Povinelli, those animals which are relatively heavy (ie. 40-80kg) encounter "qualitatively different" problems when moving from tree to tree than do those animals which are lighter. It is not made clear as to exactly what these problems are, but presumably they have something to do with the fact that there are not as many options (with respect to those parts of the tree that can be safely held) for a heavy animal as there are for a light animal. Povinelli hypothesises that the increase in body size of the great apes might have necessitated a "high-level self-representational system" in order to plan their inter- and intra-arboreal movements. This system of self-representation is a "kinesthetic self-concept". Thus self-concept is "an explicit representation of the position and movement of their own bodies".

(2) Tree Dwellers vs Ground Dwellers.

Those great apes which did not remain in the trees will presumably not have required the kinesthetic self-concept. If they are primarily ground-dwellers, then such a self-representation (which is designed to enable inter- and intra-arboreal movement) will be superfluous. Thus we can see how gorillas would not have needed to acquire such a self-concept.

It seems reasonable to assume that tree dwellers need a kinesthetic self-concept that is adequate to enable them to move through trees (although not literally of course!). The idea that this self-concept needs to be somehow better than ground-dweller's seems to me to be in need of further justification, beyond that given in the paper (I don't know - has this been done elsewhere?). Likewise, the idea that heavy animals need a greater self-concept than lighter ones also seems not uncontroversial. I take it, however, that (2) is the more controversial.

According to the logic of (2), it would seem that humans should have less of a kinesthetic concept than do orangutan's, in just the same way as gorillas. Povinelli, however, says that "[h]umans, in contrast, slowed down their growth ate, allowing more years for cognitive development". Nut why would slowing growth enable cognitive development? It seems in (1) that Povinelli is saying that it is *increased* growth, along with the issues involved with being a tree-dweller that gave rise to the kinesthetic self-concept. Now, however, he asserts that *slower* growth, combined with living on the ground lead to increased cognitive development. It is far from clear that these claims are compatible. In (1), the reason why the kinesthetic self-concept arised was because of the increase in size. In (2) it seems as though cognitive development is possible only given a lack of growth. What is going on here? Perhaps when Povinelli talks about a kinesthetic self-concept, he is talking about something different from cognitive development. But if this is so, then it would seem that humans are a counter-example to his argument. If not, then it is unclear what force his argument has.

Any suggestions?
Simon.

From bradt@U.Arizona.EDU Mon Nov 8 02:19:13 1999
Return-Path: <bradt@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.Arizona.EDU [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id CAA21167
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 8 Nov 1999 02:19:13 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JI2U28M14W9H1AAA@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Mon,
8 Nov 1999 02:20:13 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTP id <01JI2U28771C9JK70I@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 08 Nov 1999 02:20:12 -0700 (MST)
Received: from orion.U.Arizona.EDU (orion.U.Arizona.EDU [128.196.137.206])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id CAA21162 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 08 Nov 1999 02:19:04 -0700
Received: from localhost (bradt@localhost)
by orion.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id CAA14613; Mon,
08 Nov 1999 02:20:04 -0700 (MST)
Date: Mon, 08 Nov 1999 02:20:04 -0700 (MST)
From: Brad J Thompson <bradt@U.Arizona.EDU>
Subject: Re: Animal Consciousness
In-reply-to: <NCBBJCIPCKAPOGDICGPAEJOCDAA.lnielsen@azstarnet.com>
To: lnielsen@U.Arizona.EDU
Cc: scicon@paradox.soc-sci.arizona.edu, bradt@U.Arizona.EDU

Message-id: <Pine.HPX.4.10.9911080120580.10340-100000@orion.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

I agree completely with Lis' contention that Gallup's assimilation of consciousness with what he calls "self-consciousness" is confused and more importantly mistaken. If "self-consciousness" in his mouth refers to the kind of thing that the mirror test *might* test for (having a self concept etc.) then self-consciousness has nothing fundamental to do with consciousness. [I add the qualification of "fundamental" here because of course if the deployment of a self concept or a theory of mind is a conscious activity, then it will trivially have *something* to do with consciousness]. With her point against Povinelli and with her discussion of the assumptions regarding complexity, I think Lis is pointing to the fact that questions of animal consciousness are distinct from questions of the complexity of animal cognition. The question of theory of mind in animals, along with matters regarding representation and intentionality, belong in the second category.

It is easy to want to draw conclusions about the presence of consciousness from the degree of representational complexity in an animal species. I know that reading about vervet monkeys in Cheney and Seyfarth's "How Monkeys See the World" (which I highly recommend) had this effect on me, as did seeing a documentary about orangutans. It appears that we pre-theoretically do base our assessment of consciousness on representational or behavioral complexity. The question is whether this is legitimate. I think that there has to be *some* relationship here. One might start by saying that some sort of representational capacity is required for consciousness--maybe only representations are the kinds of things that can be conscious. This would at least justify not attributing consciousness to rocks. And it might justify the position that anything that is merely a behaviorist organism with no internal representations lacks consciousness. I think this might still be compatible with Lis' suggestion that crabs are conscious, depending on what constraints we place on counting as a representation. But it looks like we need something like this connection between consciousness and representation in order to avoid panpsychism.

I think we also need some kind of assumption about the relationship between consciousness and representation or cognitive complexity in order to even begin an empirical investigation of animal consciousness. I know that I'm conscious directly, and I assume that other human beings are conscious because they are physically similar to me. But what is my inroad to asking whether my cat is conscious? I could notice that she has a brain and base my answer on that--but this would be due to an assumption about what it is that makes me conscious (namely, having a brain). But you might think that having a brain makes me conscious because of what my brain *does*, and something other than a bunch of neurons could perform the same function. And one of the things that my brain does, and which seems closely connected to consciousness, is that it represents states in the environment. So maybe it is this representational capacity that we should look for in non-human animals in order to determine if they are conscious.

Perhaps increased degrees of representational complexity are correlated with different types or degrees of consciousness. This would be a very substantive claim, but if it were true it might justify looking into theory of mind as an insight into something relating to consciousness. Unfortunately I don't think this is what is going on with Gallup's

reasoning.

This leads me to what may be a point of disagreement with Lis. I would want to defend Morgan's Cannon and its use in the case of the mirror test. But this is separable from the first issue Lis points to about conflating consciousness with empathy or self-consciousness. Maybe Gallup is assuming that there is a hierarchy and that consciousness is at the top (such that "lower levels" don't involve consciousness) but this does not follow from his adherence to Morgan's Cannon alone. The placement of consciousness at the top of the purported hierarchy would be the culprit here in my opinion. But the idea of a hierarchy is not problematic, assuming that all is meant here is degree of complexity, etc. Perhaps the term "hierarchy" has moral connotations that are misleading--I would suggest abandoning the term then and just speaking of computational complexity. But it seems to me that Morgan's Cannon is appropriate for asking questions about the computational complexity of various cognitive activities in animals AND in humans. But I agree with Lis that we should be careful about making assumptions about where consciousness fits in with this matter of computational complexity.

Let me emphasize again that I think we should not attribute more complexity than necessary to explain behavior to either animals or humans. This is just Occam's Razor, as Dave mentioned. So my wanting to defend Morgan's Cannon is not the result of human-centrism. And as I suggested earlier, this is especially true given my agreement with Lis that consciousness is not some extra thing at the top of a hierarchy which only humans and maybe "higher" animals possess. Even after applying Morgan's Cannon and offering a "lower level" (in info-processing terms) explanation of an animal's behavior, we might *still* want to attribute consciousness to that animal and even to the processes in question.

Brad

Brad J Thompson
bradt@U.Arizona.EDU

From bayne@U.Arizona.EDU Mon Nov 8 08:48:57 1999
Return-Path: <bayne@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.Arizona.EDU [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id IAA21447
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 8 Nov 1999 08:48:56 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.2-31 #39830)
id <01JI37NFOXS09JKEH4@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Mon,
8 Nov 1999 08:49:56 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDf V5.2-31 #39830) with ESMTP id <01JI37NF0W809JKBPU@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 08 Nov 1999 08:49:55 -0700 (MST)
Received: from f1n7.u.arizona.edu
(IDENT:bayne@f1n7.U.Arizona.EDU [128.196.137.107])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id IAA21442 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 08 Nov 1999 08:48:42 -0700
Received: from localhost (bayne@localhost) by f1n7.u.arizona.edu (8.8.8/8.8.8)
with ESMTP id IAA44196 for <scicon@paradox.soc-sci.arizona.edu>; Mon,
08 Nov 1999 08:49:42 -0700
Date: Mon, 08 Nov 1999 08:49:42 -0700 (MST)
From: Timothy J Bayne <bayne@U.Arizona.EDU>
Subject: Re: Mirror Test
In-reply-to: <Pine.A41.4.10.9911071650370.34642-100000@f1n2.u.arizona.edu>

Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9911080843210.33890-100000@f1n7.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

If you're interested in the mirror test, you may be interested in the following papers:

(1) Robert W. Mitchell (1997) "Kinesthetic-Visual Matching and the Self-concept as Explanations of Mirror Self-Recognition" Journal for the Theory of Social Behavior 27/1, 17-39.

Mitchell's paper is a nice discussion of both Gallop and Povinelli; he rejects their accounts for a leaner model of what's going on.

(2) Lawrence Davis, (1989) "Self-consciousness in Chimps and Pigeons", Philosophical Psychology, 2/3, pp. 249-59. (U of A library doesn't have this volume.)

tim

Timothy J. Bayne
RM. 213 Social Science
Department of Philosophy
University of Arizona
Tucson, AZ 85721
USA

Hm ph. (520) 298 1930

From press@U.Arizona.EDU Mon Nov 8 19:55:52 1999
Return-Path: <press@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.Arizona.EDU [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id TAA23088
for <chalmers@paradox.soc-sci.arizona.edu>; Mon, 8 Nov 1999 19:55:52 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDF V5.2-31 #39830)
id <01JI3UYBRDJ49H1AZJ@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Mon,
8 Nov 1999 19:56:53 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDF V5.2-31 #39830) with ESMTTP id <01JI3UYBEMC09JJQOK@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Mon, 08 Nov 1999 19:56:52 -0700 (MST)
Received: from f1n3.u.arizona.edu
(IDENT:press@f1n3.U.Arizona.EDU [128.196.137.103])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTTP id TAA23083 for
<scicon@paradox.soc-sci.arizona.edu>; Mon, 08 Nov 1999 19:55:39 -0700
Received: from localhost (press@localhost) by f1n3.u.arizona.edu (8.8.8/8.8.8)
with ESMTTP id TAA51280; Mon, 08 Nov 1999 19:56:36 -0700
Date: Mon, 08 Nov 1999 19:56:36 -0700 (MST)
From: Joel K Press <press@U.Arizona.EDU>
Subject: Re: Povinelli
In-reply-to: <Pine.HPX.4.10.9911072250130.5757-100000@orion.U.Arizona.EDU>
To: Simon E Roberts-Thomson <serobert@U.Arizona.EDU>
Cc: Sarah A Wright <sawright@U.Arizona.EDU>,
scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.A41.4.10.9911081936470.38772-100000@f1n3.u.arizona.edu>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII

Status: R

All

Regarding the apparent contradiction between the human possession of a self-concept and the gorillas' apparent lack of one, it seems there is a perfectly good evolutionary explanation for this. It is true that both humans and gorillas are land dwellers rather than tree dwellers. So, by Povinelli's argument, neither has a need for a self-concept. However, gorillas diverged from humans earlier than chimpanzees, which do live in trees. So our self-concept may have developed after our gorilla cousins had already "descended from the trees." When our ancestors adopted a non-tree lifestyle, they already had a self-concept, unlike the gorillas. At first, this self-concept might not have been essential to our survival, but self-concepts are generally useful, so as long as the cost of keeping it isn't too high, it probably won't be bread out. In fact, Povinelli would probably argue that our purportedly more robust sense of self evolved out of the old one because it turned out to be the case that self-concepts are valuable in a wide variety of environments.

Joel

From switanek@U.Arizona.EDU Tue Nov 9 01:51:33 1999
Return-Path: <switanek@U.Arizona.EDU>
Received: from Arizona.EDU (Maggie.Telcom.Arizona.EDU [128.196.128.233])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id BAA23811
for <chalmers@paradox.soc-sci.arizona.edu>; Tue, 9 Nov 1999 01:51:32 -0700
Received: from DIRECTORY-DAEMON by Telcom.Arizona.EDU (PMDf V5.2-31 #39830)
id <01JI47DBEQDC9H0DW0@Telcom.Arizona.EDU> for
chalmers@paradox.soc-sci.arizona.edu (ORCPT rfc822;chalmers@arizona.edu); Tue,
9 Nov 1999 01:52:34 MST
Received: from paradox.soc-sci.arizona.edu by Telcom.Arizona.EDU
(PMDf V5.2-31 #39830) with ESMTP id <01JI47DB2MKG9JJYRC@Telcom.Arizona.EDU>
for chalmers@Arizona.EDU; Tue, 09 Nov 1999 01:52:33 -0700 (MST)
Received: from pavo.U.Arizona.EDU (pavo-2.U.Arizona.EDU [128.196.137.196])
by paradox.soc-sci.arizona.edu (8.9.3/8.9.3) with ESMTP id BAA23806 for
<scicon@paradox.soc-sci.arizona.edu>; Tue, 09 Nov 1999 01:51:02 -0700
Received: from localhost (switanek@localhost)
by pavo.U.Arizona.EDU (8.8.6 (PHNE_17190)/8.8.6) with ESMTP id BAA02577; Tue,
09 Nov 1999 01:52:03 -0700 (MST)
Date: Tue, 09 Nov 1999 01:52:03 -0700 (MST)
From: Nicholas J Switanek <switanek@U.Arizona.EDU>
Subject: Re: Animal Consciousness and Nagel's answer
In-reply-to: <NCBBJCIPCKAPOGDICDGP AEJOCDA A.lnielsen@azstarnet.com>
To: lnielsen@U.Arizona.EDU
Cc: scicon@paradox.soc-sci.arizona.edu
Message-id: <Pine.HPX.4.10.9911090050010.26598-100000@pavo.U.Arizona.EDU>
MIME-version: 1.0
Content-type: TEXT/PLAIN; charset=US-ASCII
Status: R

Lis' suggestion not to be too anthropocentric reminded me of a wonderful idea that was discussed in class last Tuesday. It was that maybe gorillas notice the spot, but just don't get bothered about it. We seem most able to empathize and then impute consciousness to those species that share our vanities. I, too, will become agitated when my teeth start falling out, and would freak out if someone dyed my hair pink while I was sleeping. If a tell-tale sign of self-consciousness is vanity (and we might recall the pre-university, round about middle- to high-school, use of the word 'self-conscious'), then indeed we might be embarrassed to put self-consciousness at the top of the hierarchy of cognitive complexity.

I wonder not about chimps learning, which is generally considered not to require consciousness, but about chimps teaching. Might not chimps have to empathize with their pupils, believe that the pupils' worldviews are rather akin to their own? Otherwise, it seems chimps couldn't be all that hopeful of success in teaching.

I still find Nagel's article to be rather defeatist. First of all there exists an essential otherness to another subject's point of view that prevents me from ever fully comprehending that subject's experience, at least as far as imagination might be a means of getting closer to what it is like for a subject to experience something as that subject. And Nagel thinks this, imagination or "adopting [the subject's] point of view", is most natural way we go about answering the question What is it like...? Imagination gets us part of the way because it is something like experience, which is the means by which we know "what it is like to be us." Indeed, in trying to access the phenomenality of a bat, "The best evidence would come from the experience of bats, if we only knew what they were like." Experience or the imagination of experience, and not explanation, is the most common way of answering What is it like? And although, because in part of a recognition of the constraints on our capacities for experience and imagination, we realize we must turn to explanations in order to answer the question in the case of other subjects, especially those from other species, still we are frustrated by the unfeeling answers of a primitive science. So when we measure them against our expectations of a true and robust answer, we inevitably find objective answers lacking. "If the subjective character of experience is fully comprehensible only from one point of view, then any shift to greater objectivity--that is, less attachment to a specific viewpoint--does not take us nearer to the real nature of the phenomenon: it takes us farther away from it."

This last quote of Nagel's is bleak. Moving towards a more objective perspective is wrongheaded, and gaining the subject's point of view is impossible. I think Nagel recognizes that a move towards the objective is wrongheaded only insofar as such a move is presently counterintuitive. But our present intuition is that the only satisfactory answer to What is it like...? is an experience or imagination thereof. If a science of consciousness could explain "_how_" the physical and the mental are linked, maybe one could learn to accept physicalist answers, which should in turn allow robust answers to Nagel's title question, and others like it. Akins' article demonstrates the substantial success one can have in moving towards the objective, and incidentally, as we explored in class, give us a much more accurate idea of what we might try to imagine.

But I did have one problem with Akins' article. How are decibels measured? The mustached bat emits beeps and swooping screeches at 100 decibels, "as loud as a rock concert." The jungle must be a terrifying place at night! But how is this supposed to work, since a rock concert can be heard for miles, but a bat can only count on its music carrying a measly six feet?

nick

INDV 101 - PHILOSOPHICAL PERSPECTIVES ON THE INDIVIDUAL - PAPER TOPICS

SECOND PAPER

Due at the beginning of class (2pm), Wednesday December 1, 1999. This should be a 4-5 page typed, doubled-spaced paper. Include your name and your TA's name on the front. Late papers will be penalized at a rate of one letter-grade per day.

You have the option of submitting a rough draft of your paper to your TA by 2pm on Monday November 22. Your TA will return it to you with comments on Monday November 29, and the final draft will be due at the normal time. Assessment is based on the final draft. This process is optional but recommended, both to improve your paper and to improve your grade.

For instructions on how to write a philosophy paper, read the following carefully!

- [Writing a Philosophy Paper](#) (Peter Horban)
- [Guide to the Study of Philosophy](#) (Garth Kemerling)
- [What Does He Want?](#) (Geoffrey Payzant)
- [Undergraduate Philosophical Writing](#) (Martin Young)
- [Guidelines on Writing a Philosophy Paper](#) (James Pryor)

1. Present and explain Jackson's argument against materialism. Critically assess the argument. Is it valid? Are the premises true? What would an intelligent opponent (say, Jackson if you are arguing against him, or a smart materialist if you agree with Jackson) say to rebut your analysis? Present their case as strongly as possible, then rebut it.

2. Present and explain Chalmers' distinction between the "easy" and the "hard" problems of consciousness, and his argument that no purely physical approach can solve the hard problem. Critically assess the argument. Does it establish its conclusion? What would an intelligent opponent say to rebut your analysis? Present their case as strongly as possible, then rebut it.

3. Present and explain either the "perfect actor" or "super-Spartan" argument against behaviorism. Put the argument in the form of a formal argument, and then critically assess it. Is it valid? Are the premises true? What would an intelligent opponent say to rebut your analysis? Present their case as strongly as possible, then rebut it.

4. An argument against free will runs as follows: (1) If determinism is true, we have no free will, (2) Determinism is true, so (3) We have no free will. Assess the argument. If you do not accept the conclusion, which premise do you reject, and why? Depending on your choice, defend your position by answering the following questions, with reasons:

If you reject premise 1: How can free will be compatible with determinism? Isn't a deterministic life like a boat that is steering down a single underlying track with no freedom? How can we make sense of moral responsibility and choice if all actions are predetermined?

If you reject premise 2: How can nondeterminism support free will any better? If decisions are determined by random dice-rolling, is this really something that we have control over? And if we control the dice-rolling, isn't that determinism after all?

If you accept the conclusion: If there is no free will, how can there be moral responsibility? What is the difference between actions we are responsible for and actions we are not responsible for, if it is not the fact that we chose freely? If there is no freedom or moral responsibility, is it fair to punish people for acts for which they were not responsible?

FIRST PAPER

Due at the beginning of class (2pm), Monday October 4, 1999. This should be a 4-5 page typed, double-spaced paper. Include your name and your TA's name on the front. Late papers will be penalized at a rate of one letter-grade per day.

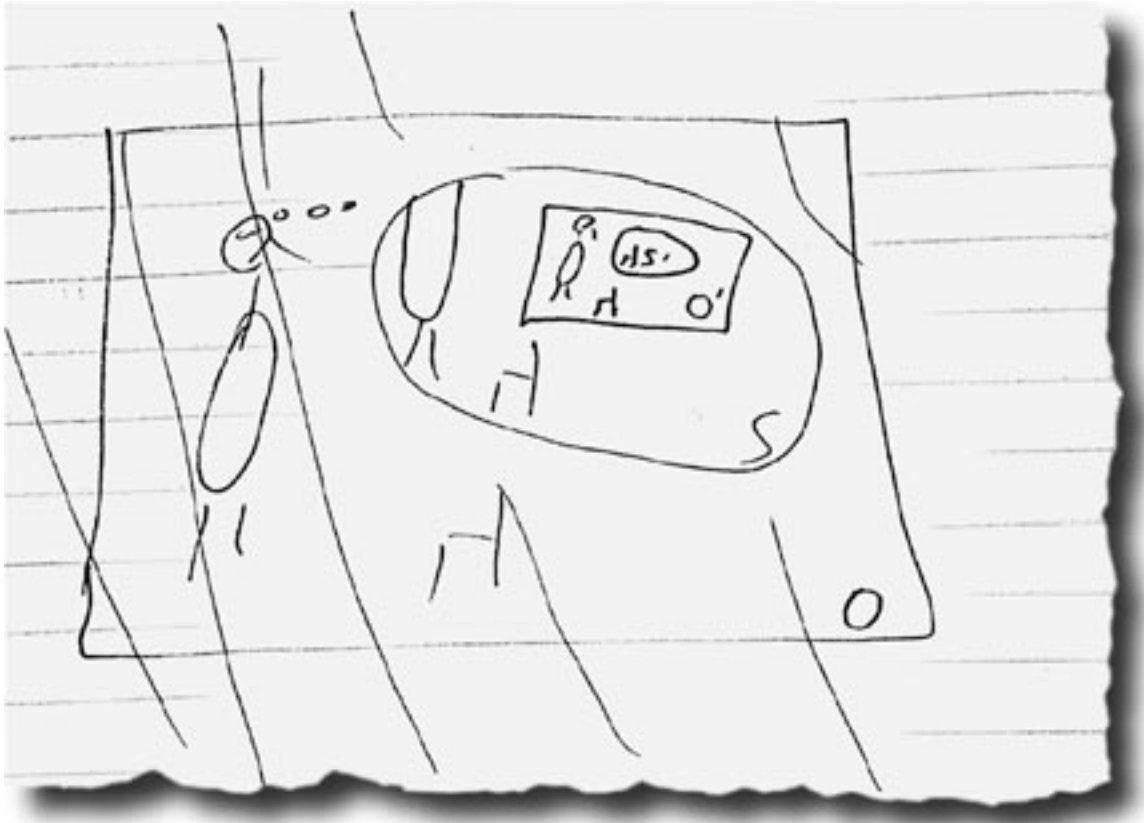
1. Present and explain one of Descartes' arguments for his version of dualism. Critically assess the argument. Is it valid? Are the premises true? What would an intelligent opponent (say, Descartes if you are arguing against him, or a smart materialist if you agree with Descartes) say to rebut your analysis? Present their case as strongly as possible, then rebut it.
2. Smullyan's piece "An Unfortunate Dualist" is intended to cast doubt on dualism. Set out explicitly the argument that you think is underlying Smullyan's reasoning. Critically assess this argument (validity and premises). How would a (smart) interactionist respond? How would a (smart) epiphenomenalist respond? What do you think the upshot is?
3. Present Descartes' skeptical challenge (for the conclusion that we cannot know about external reality) in the form of an argument. Critically assess this argument (validity and premises). If you think the argument fails and that you can know that external reality exists, say how you think this knowledge is possible. If you think the argument succeeds, what do you think the upshot is? Either way: what would an intelligent opponent say to rebut your analysis, and how would you respond?
4. You are Keanu Reeves (Neo) in "The Matrix". Laurence Fishburne (Morpheus) finds you and tells you that you are in the Matrix, and says he will help you return to the real world. Is there any way for you to truly know that you are in the Matrix to start with, or that you are in reality later? If yes, could the movie situation be modified so that you could not know these things? If you cannot truly know, is it rational for you to believe that you are in the Matrix at the beginning or that you are in the reality later? Give reasons in every case, present at least some aspect of your reasoning in the form of a formal argument, and

answer objections from an intelligent opponent who would disagree with you. What does this situation teach us about knowledge and external reality, and why?

Bonus question: (no formal credit for this, but fun): Are there any philosophical flaws in the movie "The Matrix"? Give reasons.

[INDV 101 home page](#)

Consciousness has scientists of all kinds scheming. Dr. Piet Hut, a physicist, offers this sketch: the objective world (o), with person and chair; the subjective experience of the world (s), in the balloon; the empirical understanding of the world, including physics (o prime), in the smaller rectangle; and the empirical understanding of subjective experience, including psychology, (s prime), in the really small balloon. Dr. Hut argues that when s and s prime are conflated, confusion reigns.



[Go to Article](#)

Copyright 1997 The New York Times Company